Semi-Nonparametric Indirect Inference

Blasques, F.

2011

Semi-Nonparametric Indirect Inference

Francisco Blasques

This book was typeset by the author using LaTeX.

Published by Universitaire Pers Maastricht
ISBN: 978-94-6159-091-6
Printed in The Netherlands by Datawyse Maastricht

Semi-Nonparametric Indirect Inference

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit Maastricht,
op gezag van de Rector Magnificus,
Prof. mr. G.P.M.F. Mols,
volgens het besluit van het College van Decanen,
in het openbaar te verdedigen
op donderdag 3 november 2011 om 12.00 uur

door

Francisco Albergaria Amaral Blasques

UNIVERSITAIRE
PERS MAASTRICHT

Promotores:

Prof. dr. B. Candelon

Prof. dr. J.R.Y.J. Urbain

Co-Promotores:

Dr. E. Beutner

Beoordelingscommissie:

Prof. dr. F.C. Palm (voorzitter)

Prof. dr. A. Lucas (Vrije Universiteit Amsterdam)

Prof. dr. A. Monfort

Dit onderzoek werd financieel mogelijk gemaakt door de Maastricht Research School of Economics of Technology and Organizations (METEOR).

Acknowledgements

Several friends and colleagues have helped me in writing this thesis. Most importantly, they have made this task enjoyable.

I am very grateful to my supervisors Eric Beutner, Bertrand Candelon and Jean-Pierre Urbain. Not only was their guidance and constant support essential for my work, their friendship was a great source of happiness during this period.

I would like to thank the members of the reading and assessment committee, Andre Lucas, Alain Monfort and Franz Palm, for their careful reading of the manuscript and for the many insightful comments and suggestions.

I am thankful to Karin van den Boorn and Haydeé Hallmanns for their help and assistance. I am also thankful to Bram Driesen for his help in preparing the Nederlandse samenvatting.

To all my colleagues and friends in Maastricht and elsewhere my sincere thanks and gratitude. The names are too many to mention. You know who you are.

Finally, I am most thankful to my family. My wonderful brother Ze Pedro, my dear mother Elisabete and my lovely Rita. I dedicate this thesis to the loving memories of my father, Jose Blasques.

Francisco Blasques
Maastricht, 2011.

Contents

Chapter 1

Introduction

This thesis proposes a solution to a couple of problems that afflict the statistical analysis of economic data. These problems are in some sense 'opposite' to each other. Loosely speaking, the first problem is related to *excessive model simplicity*. The second is related to *excessive model complexity*.

The problem of *excessive model simplicity* is well known in econometric analysis. The use of models that are 'excessively simple' will typically result in some form of *model misspecification*. This might be problematic since econometric analysis rests frequently on axioms of correct specification whose reasonability is, at the very least, questionable.

The problem of *excessive model complexity* is common in the estimation of non-linear dynamic models derived from economic theory. The specification of 'complex' dynamic models often results in difficulties with classical estimators (e.g. when criterion functions become analytically intractable).

To tackle both problems, the solution proposed in this thesis involves the use of an extremum sieve estimator for semi-nonparametric models that relies on auxiliary statistics through the principle of indirect inference. To settle ideas, I propose that we start immediately by looking at a couple of simple examples. Hopefully, these will illustrate the need for an estimation methodology that deals simultaneously with both *excessive model simplicity* and *excessive model complexity*.

Excessive Simplicity

Let (y_1, x_1), (y_2, x_2), ... be a random sample from the joint distribution of y and x. Suppose that we are interested in conducting statistical inference on the conditional expectation function of y given x, denoted by $\boldsymbol{\theta}_0 \equiv E(y|x)$. In general, it can be said that our objective amounts to 'searching' for $\boldsymbol{\theta}_0$ on a space Θ of possible 'candidates' for the conditional expectation $E(y|x)$. Naturally, this leads us to the formulation

of a regression model of the type,

$$y_t = \boldsymbol{\theta}(x_t) + \epsilon_t$$

where $\boldsymbol{\theta} \in \Theta$ is a 'candidate' for the conditional expectation $E(y|x)$ and ϵ_t is accordingly defined as $\epsilon_t := y_t - \boldsymbol{\theta}(x_t)$. The problem of excessive model simplicity is well known in econometric analysis and it usually comes under the label of *model misspecification*.

One important source of misspecification resides in the choice of Θ, the space on which we 'search' for the function $\boldsymbol{\theta}_0$. If this space is 'too small', then it might happen that $\boldsymbol{\theta}_0 \notin \Theta$. In other words, the model becomes *misspecified*. As we shall see, model misspecification is not a statistical problem in itself since most estimators $\hat{\boldsymbol{\theta}}_T$ can still be shown to converge to a limit $\boldsymbol{\theta}_0^* \in \Theta$ that might possess interesting properties. However, it does pose important problems to the *econometric* analysis of statistical results.

In an effort to avoid this problem, one might thus be tempted to define a space Θ that is as large as possible. Unfortunately, in such spaces, most estimators will fail to be of any use. This problem was elegantly formulated by Geman and Hwang (1982) in a regression example just like the one formulated above.

Remark 1.0.1. *Suppose that we let Θ be the entire space of continuous functions. Then for every sample size T, there exists (almost surely) a set of functions $\Theta^* \subset \Theta$ that 'passes through' all the sample points (y_1, x_1), (y_2, x_2), ... Every function $\boldsymbol{\theta}$ in the set Θ^* yields a 'perfect fit' and 'maximizes likelihoods'. However, the set Θ^* of optima does not converge in any meaningful way to $\boldsymbol{\theta}_0$. In essence, the problem of model misspecification (or excess simplicity) has not been solved. On the contrary, the generality of Θ has gave way to a failure of statistical estimators to be 'consistent' to $\boldsymbol{\theta}_0$ and thus to provide any meaningful 'information' about the parameter of interest.*

Fortunately, a solution to this problem has been proposed by Grenander (1981) and it goes by the name of *method of sieves*. Grenander's method of sieves proceeds by restricting the estimator $\hat{\boldsymbol{\theta}}_T$ to take values in specially chosen subsets of Θ that still allow for $\hat{\boldsymbol{\theta}}_T$ to be 'consistent' for any $\boldsymbol{\theta}_0 \in \Theta$. The method of sieves is fundamentally related to the *semi-nonparametric* modeling approach and it will be an integral part of the work contained in this thesis. For now, let us just keep in mind that the method of sieves is in effect capable of dealing with large parameter spaces Θ, thus offering a more convincing solution to the problem of *model misspecification* or *excessive simplicity*.

> *"The method of sieves leads easily to consistent nonparametric estimators in even the most general settings."* in Geman and Hwang (1982)

Excessive Complexity

Suppose now that the parameter of interest $\boldsymbol{\theta}_0 \equiv E(\mathbf{x}_t|\mathbf{x}_{t-1})$ defines a conditional expectation in the context of a possibly nonlinear vector autoregressive model. In particular, consider a vector process $\{\mathbf{x}_t,\, t \in \mathbb{Z}\}$, containing both observed and latent variables, whose distribution is implicitly defined by the dynamic equation,

$$\mathbf{x}_t = \boldsymbol{\theta}_0(\mathbf{x}_{t-1}) + \boldsymbol{\epsilon}_t$$

where $\{\boldsymbol{\epsilon}_t,\, t \in \mathbb{Z}\}$ are innovations with common distribution $\mathrm{D}_{\boldsymbol{\epsilon}}$. In such settings (especially when $\boldsymbol{\theta}_0$ is nonlinear) classical estimation procedures might fail to be useful. For example, likelihood functions might be intractable, yielding maximum likelihood (ML) techniques inappropriate. Similar problems occur when moment conditions can not be derived analytically, yielding method of moments (MM) estimators equally inappropriate. Loosely speaking, this problem is what we refer to as the problem of *excessive model complexity*. In essence, it is the 'complexity of the model' that prevents us from deriving classical estimators for the parameter of interest.

Again, we are fortunate enough to have a solution: *simulation-based estimation procedures*. The statistical inferential principle underlying most simulation-based estimation methods goes by the name of *indirect inference*. This unifying principle was introduced in Gouriéroux and Monfort (1993); see also Smith (1993). Very simply, our objective in this thesis will consist of combining the *method of sieves* with the *principle of indirect inference* to produce an estimation method capable of dealing simultaneously with the problems of *excessive simplicity* and *excessive complexity* of models in econometrics.

> *"... indirect inference [...] allows for estimation and test procedures when the model is too complicated to be treated by usual methods."*
> in Gouriéroux and Monfort (1993)

Prevalence of the Problem

It is somewhat ironic that we should experience problems related to excessive model complexity, when (at the same time) we worry about the model's excessive simplicity, and search for methods that give us greater generality (i.e. methods capable of describing data generating processes of greater complexity). Yet, this is quite a natural state of affairs in econometrics. In general, economic theory suggests that economic variables are often related in a complex dynamic nonlinear fashion. This is true for most models ranging from applied macroeconomics to microeconomics and empirical finance. The natural implication of this is that models derived from theory are typically challenging to estimate.

> *"It seems to be generally accepted that the economy is nonlinear, in that major economic variables have nonlinear relationships. Economic theorists suggest models with floors and ceilings, buffer stocks, and switching regimes. Investment functions, production functions, and Phillips' curves are usually specified in nonlinear forms."*
> in Granger and Terasvirta (1993)

Regardless of the complexity suggested by theory. Probability models derived from theoretical postulates are at the same time too simplistic for axioms of correct specification to hold with any reasonable degree of confidence. Numerous accounts of this (almost inherent) feature of econometric modeling could be given here based solely on empirical evidence. The prevalence of model misspecification problems in econometrics is indeed generally recognized. It is thus not surprising to find the above quote being followed almost immediately by the following remark.

> *"However, most economic theories only suggest plausible nonlinear relationships, usually are incomplete, and often do not agree with actual data, particularly in the dynamic structure. There thus seems to be a need for exploratory statistical techniques to produce sound models, perhaps used in conjunction with appropriate theories."*
> in Granger and Terasvirta (1993)

It is my belief that the method of sieves and semi-nonparametric models embody quite well the 'exploratory' spirit of econometric analysis called for in this quote.

The Structure of this Thesis

In the rest of this chapter, the interested reader will find a brief account of the literature of approximation theory, extremum estimators, sieve estimators, Semi-NonParametric (SNP) models and the method of indirect inference. In particular, the pages below offer some historical background on several statistical developments that support the theory of Semi-Nonparametric Indirect Inference (SNPII). This body of literature is immensely vast, and there is no pretension of providing a survey that is even remotely exhaustive or complete. Here I will simply lay down the developments that seem most relevant for the theory that follows.

Section 1.1 begins with a review of the literature of extremum estimators and its classical proof of consistency. Section 1.2 extends the discussion to the convergence rate and asymptotic distribution of extremum estimators on infinite dimensional spaces. Section 1.3 introduces the method of sieves and SNP models. The theory of sieve estimators and SNP models turns out to rely fundamentally on concepts of function approximation. As a result, Section 1.4 summarizes the rich history of

Approximation Theory. Section 1.5 introduces the principle of indirect inference and its relation to simulation based estimators. Finally, Section 1.6 gives a first superficial introduction to the idea of semi-nonparametric indirect inference, which combines the (until now) separate strands of literature of *SNP models* and *indirect inference*. A motivating econometric example is used to place the SNPII estimator in context with the literature covered in Sections 1.1-1.5. In the remaining chapters of this thesis we study more carefully the SNPII estimator.

Chapter 2 introduces the novel SNPII estimator in its most general form and provides a first account of its properties. In particular, this chapter delivers the main results of consistency, convergence rate and asymptotic distribution of the SNPII estimator. While the consistency of the SNPII estimator is obtained as a special case of existing theorems for sieve extremum estimators, the convergence rate and asymptotic distribution theorems are entirely new. These two theorems apply to a large class of smooth sieve estimators and thus constitute an addition to the general theory of sieve extremum estimation.

Chapter 3 provides a more rigorous treatment of the SNPII theory. This is done in the context of an SNPII estimator that relies on an infinite number of parametric auxiliary statistics. The estimator is shown to be $\sqrt{T}$ consistent and asymptotically Gaussian under general regularity conditions. The data is allowed to exhibit heterogeneous and dependent behavior. Furthermore, in the tradition of indirect inference, these results apply to a large class of complex dynamic models with unobserved variables. In particular, including those yielding an estimator with no closed form algebraic representation or featuring a criterion function which is intractable or infeasible, even on appropriately chosen compact finite-dimensional sieves. These results add to the theory of sieve estimation which implicitly assume analytical tractability and typically impose considerably more restrictive conditions on data dependence and heterogeneity.

Chapter 4 provides some first Monte Carlo evidence of the finite-sample behavior of the SNPII estimator in a couple simple settings. The evidence gathered in this chapter seems to confirm the theoretical results of Chapters 2 and 3.

Proofs of consistency of extremum estimators usually require assumptions ensuring that there exists a unique well separated (*identifiably unique*) minimizer of the limit criterion function. Unfortunately, these assumptions are sometimes opaque and do not lend themselves to immediate verification. Chapter 5 provides methods for confirming that *identifiable uniqueness* holds for the class of extremum estimators whose limiting criterion function can be appropriately defined as a divergence on a space of probability measures.

Chapter 6 summarizes the main findings of this thesis and concludes.

1.1 Extremum Estimators

An extremum estimator, denoted $\hat{\boldsymbol{\theta}}_T$, is typically defined as the minimizer (or maximizer) of a random *criterion function* Q_T on a parameter space Θ,[1]

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\boldsymbol{\theta}). \tag{1.1}$$

The criterion Q_T is random because it is a function of random variables $X_1, ..., X_T$. We could thus have written $Q(X_1, ..., X_T, \boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}(X_1, ..., X_T)$ instead of $Q_T(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_T$ respectively. For simplicity however, we adopt the shorter notation used in (1.1).

In general, we are interested in showing that the extremum estimator $\hat{\boldsymbol{\theta}}_T$ converges in an appropriate fashion to $\boldsymbol{\theta}_0$, defined to be the minimizer of the limit deterministic criterion function Q_∞,

$$\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}).$$

Based on ideas that dated back to the work of Doob (1934, 1953), Cramer (1946), Wald (1949) and Le Cam (1953) on the consistency of the *Maximum Likelihood* (ML) estimator, the classical proof of consistency of extremum estimators was first laid down in Jennrich (1969) and Malinvaud (1970).

Remark 1.1.1. *Proofs of consistency existed already for various parametric extremum estimators dealing with linear models. Valuable developments in econometrics included e.g. the theory of Maximum Likelihood (ML) and Least Squares (LS) estimation of autoregressive models; see e.g. Mann and Wald (1943).*

By establishing the consistency of LS estimators in nonlinear regression models with fixed regressors and *independent identically distributed* (iid) residuals, the work of Jennrich (1969) and Malinvaud (1970) marked an important step in the development of a general consistency theory for extremum estimation of nonlinear models. Subsequent developments included *(i)* extensions to multivariate regression settings, *(ii)* allowing for stochastic regressors and *(iii)* the weakening of the iid residuals assumption; see e.g. Hannan (1970), Robinson (1972), Gallant (1975), White (1980a) and Wu (1981).[2] In its present form, the proof of consistency of extremum estimators is typically presented in the following very elegant way (see e.g. Pötscher and Prucha 1997).

[1] If a minimum of Q_T does not exist, $\hat{\boldsymbol{\theta}}_T$ can alternatively be defined as $\hat{\boldsymbol{\theta}}_T = \inf_{\boldsymbol{\theta}\in\Theta} Q_T(\boldsymbol{\theta})$. If several minima exist, then $\hat{\boldsymbol{\theta}}_T$ can be defined as an element of the $\arg\min$ set, i.e. $\hat{\boldsymbol{\theta}}_T \in \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\boldsymbol{\theta})$. More generally, $\hat{\boldsymbol{\theta}}_T$ must satisfy $\hat{\boldsymbol{\theta}}_T \in \inf_{\boldsymbol{\theta}\in\Theta} Q_T(\boldsymbol{\theta}) + O_p(\delta_T)$ with $\delta_T \to 0$ as $T \to \infty$.

[2] The work of Gallant (1975) already embodied the spirit of semi-nonparametric modeling that was to be developed later.

Lemma 1.1.1. (Consistency of Extremum Estimator) *Let $\hat{\boldsymbol{\theta}}_T$ be defined according to (1.1) where Θ is a compact set. Let $\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0$. Finally, suppose that Q_∞ is continuous on Θ and has a unique minimizer $\boldsymbol{\theta}_0$. Then, $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$ as $T \to \infty$.*[3]

The continuity of Q_∞ is designed to ensure that $\boldsymbol{\theta}_0$ is a *well separated* or *identifiably unique* minimizer of Q_∞ on the compact Θ. Chapter 5 discusses the *identifiable uniqueness* of $\boldsymbol{\theta}_0$ in more detail and provides alternative conditions for it to hold.

Aside the uniqueness condition, it is clear that the convergence of minimizers $\{\hat{\boldsymbol{\theta}}_T\}$ to $\boldsymbol{\theta}_0$ boils down essentially to showing that Q_T converges uniformly to the limit criterion Q_∞. Often this is obtained by applying a *Uniform Law of Large Numbers* (ULLN) to the sequence of criterion functions $\{Q_T\}$ when each Q_T is given by,

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\theta\in\Theta} Q_T(\boldsymbol{\theta}) \equiv \arg\min_{\theta\in\Theta} Q_T(X_1, ..., X_T, \boldsymbol{\theta}) \equiv \arg\min_{\theta\in\Theta} 1/T \sum_{t=1}^{T} \mathcal{Q}(X_t, \boldsymbol{\theta}). \tag{1.2}$$

The estimator $\hat{\boldsymbol{\theta}}_T$ in (1.2) is called an *M-estimator*. An M-estimator is a simple generalization of the usual ML and least-squares estimators and a special case of the extremum estimator in (1.1).

Remark 1.1.2. *In M-estimation theory, consistency results boil down essentially to showing that a ULLN applies to the sequence of criterion functions $\{Q_T\}$. Likewise, asymptotic distribution results rely on showing that a* Central Limit Theorem *(CLT) holds for the real sequence $\{\sqrt{T}\partial Q_T(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\}$.*

Extensions to the theory of Jennrich (1969) and Malinvaud (1970) that relaxed the iid assumption and allowed for time dependence followed some time after. Since both ULLNs and CLTs were available for weak and strong mixing sequences, the theory of extremum estimation expanded naturally in that direction; see e.g. Domowitz and White (1982), White and Domowitz (1984), Bates and White (1985), Burguete et al. (1982), and Domowitz (1985). The use of weak and strong mixing sequences was however quite unsatisfactory in the context of dynamic models as these forms of 'fading memory' are not preserved by transformations involving the infinite past of random variables; see e.g. Andrews (1984). Results based on notions of 'fading memory' that are appropriate for dynamic models (e.g. *near epoch dependence*) followed in the important contributions of Gallant (1986) and Gallant and White (1988b). These authors used a result in McLeish (1975) to conclude that *near epoch dependent* processes are mixingales, and hence, that these satisfied LLNs and CLTs; see also Pötscher and Prucha (1991a,b).

[3]Note that $\xrightarrow{p}$ denotes convergence in probability. Note also that almost sure convergence of $\hat{\boldsymbol{\theta}}_T$ (denoted $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$) is obtained when $\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{a.s.} 0$.

Finally, as the theory of ULLNs evolved, new results became available under various forms of heterogeneity and dependence. A particularly important development consisted of the appearance of *generic ULLNs* and *generic uniform convergence* results by the hand of Newey (1991), Andrews (1987), Andrews (1992) and Potscher and Prucha (1989, 1994)). These results reduced the verification of uniform convergence to that of pointwise (or local) convergence plus some stochastic equicontinuity condition.[4] The following lemma sketches a typical *generic uniform convergence* theorem (see e.g. Davidson 1994, p.337).

Lemma 1.1.2. (Generic Uniform Convergence) *Let* (Θ, δ_Θ) *be a totally bounded metric space and* $\{Q_T\}$ *be a sequence of random real-valued functions on* Θ *satisfying* $Q_T(\boldsymbol{\theta}) \xrightarrow{p} Q_\infty(\boldsymbol{\theta})$ *for every* $\boldsymbol{\theta} \in \Theta$. *Furthermore, suppose that the sequence* $\{\Delta Q_T\}$ *with* $\Delta Q_T := Q_T - Q_\infty$ *is asymptotically uniformly stochastically equicontinuous on* Θ, *i.e. suppose that for every* $\epsilon' > 0$, $\exists\, \epsilon > 0$ *such that,*[5]

$$\limsup_{T\to\infty} \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta} \sup_{\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|<\epsilon} \Big|\Delta Q_T(\boldsymbol{\theta}') - \Delta Q_T(\boldsymbol{\theta})\Big| \geq \epsilon' \Big) < \epsilon'.$$

Then, $\sup_{\boldsymbol{\theta}\in\Theta} \Big|Q_T(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})\Big| \xrightarrow{p} 0$ *as* $T \to \infty$.

The generality of the above result is quite remarkable. When, Q_T takes the form of a sample average, then it allows us to conclude that $\{Q_T\}$ converges uniformly to Q_∞ on Θ as long an LLN applies to $\{Q_T(\boldsymbol{\theta})\}$ for every $\boldsymbol{\theta} \in \Theta$ and $\{Q_T - Q_\infty\}$ satisfies an appropriate equicontinuity condition. As we shall see, the requirement of a totally bounded metric space is however quite restrictive when considering infinite dimensional parameter spaces.

1.2 Extensions to Infinite Dimensional Spaces

The theory of extremum estimators above made no references to assumptions on *true parameters* or *correct specification*. Indeed, consistency of $\hat{\boldsymbol{\theta}}_T$ towards $\boldsymbol{\theta}_0$ does not require us to make any statements about the role that $\boldsymbol{\theta}_0$ plays in relation to the *data generating process* (DGP). Essentially, $\boldsymbol{\theta}_0$ needs only be the unique minimizer of Q_∞.[6] Often however, it is desirable to relate $\boldsymbol{\theta}_0$ to the underlying DGP, in which case the concept of model misspecification become relevant.

[4] Several important extensions came also from the field of *Empirical Process Theory* which delivered the ability to obtain uniform convergence results by controlling essentially the complexity of the class of functions over which uniformity is required. See e.g. Andrews (1986), Pollard (1989, 1990) and van der Vaart (1995).

[5] Note that $S(\boldsymbol{\theta}_0, \epsilon)$ denotes a ball of radius $\epsilon > 0$ centered at $\boldsymbol{\theta}_0$.

[6] The notion of *true parameter* and *model misspecification* exists only in relation to a DGP, i.e. in relation to a distribution or probability measure "from which the data is drawn".

The literature on model misspecification is quite vast and encompasses a large number of fields of research.[7] Statistically however, the presence of model misspecification is not always a concern. Indeed, many estimators can be shown to converge to a limit possessing interesting properties. For example, ML estimators can be shown to converge to so-called *pseudo-true parameters* having important information theoretic properties; see e.g. Akaike (1973), Akaike (1981), White (1982) and Gourieroux et al. (1984).[8] Results like these suggest quite naturally a reappraisal of the role of misspecified models in econometrics; see Monfort (1996).

Despite the possibility of conducting valuable statistical inference under model misspecification, econometric analysis still relies quite dramatically on the satisfaction of axioms of correct specification, whose reasonability is questionable. Most importantly, econometric studies are often supposed to give an answer to questions that can only take place in a world of correctly specified models. This is most clear when taking to the data models derived from economic theory.

Remark 1.2.1. *In economic theoretic models, parameters have a well defined economic meaning. In econometrics, such parameters can be estimated only under the influence of an axiom of correct specification. In the presence of an incorrectly specified model, estimators might still be consistent to some 'pseudo-true' parameter of interest. However, pseudo-true parameters are generally non-unique (see Chapter 5) and do not possess the 'deep economic meaning' proposed by economic theory.*

The remark above suggests that econometric techniques that search for generality and attempt to make correct specification axioms less restrictive have an important role to play. Here, we wish to focus on avoiding the restrictiveness of finite dimensional parameter spaces.

Let Θ be an infinite dimensional vector space. Then, Θ has (by definition) an infinite number of basis vectors that span it. Clearly, any sufficiently smooth function defined on Θ (such as Q_T and Q_∞) will have an infinite number of partial derivatives; i.e. an infinite number derivatives in the direction of its basis vectors.[9] Let us denote the system of partial derivatives of Q_T and Q_∞ by ∇Q_T and ∇Q_∞ respectively. Following van der Vaart (1995) and van der Vaart and Wellner (1996, Chapter 3.3), we give ∇Q_T the interpretation of an infinite system of estimating equations. It is important to notice that $\nabla Q_T(\boldsymbol{\theta})$ denotes the random vector of all partial derivatives of Q_T at $\boldsymbol{\theta}$, i.e. a random element in $\mathbb{R}^\infty$ (assuming *countably* infinite dimensions).[10] Accordingly, $\nabla Q_\infty(\boldsymbol{\theta})$ is a point in $\mathbb{R}^\infty \; \forall \boldsymbol{\theta} \in \Theta$.

[7] For example, the field of robust statistics originated in the contributions of Huber (1967, 1974).

[8] The ML estimator is the minimizer of the divergence introduced in Kullback and Leibler (1951).

[9] In infinite dimensional spaces, various notions of differentiability can be devised. For the moment, let us abstract from these considerations and focus solely on the argument.

[10] $\mathbb{R}^\infty$ denotes the Cartesian product of infinite copies of $\mathbb{R}$ (the set of real numbers).

Remark 1.2.2. *Under appropriate regularity conditions $\hat{\boldsymbol{\theta}}_T$ admits a Z-estimator formulation as a random variable satisfying $\nabla Q_T(\hat{\boldsymbol{\theta}}_T) = 0$ with $\nabla Q_\infty(\boldsymbol{\theta}_0) = 0$.*[11]

It is with this Z-estimator formulation of $\hat{\boldsymbol{\theta}}_T$ that we shall proceed to derive important convergence results for our extremum estimator on an infinite dimensional space. The following lemma is from van der Vaart (1995) and van der Vaart and Wellner (1996).

Lemma 1.2.1. (Convergence Rate and Asymptotic Distribution) *Let Q_T and Q_∞ be differentiable on Θ and ∇Q_∞ be continuously differentiable on an neighborhood of $\boldsymbol{\theta}_0$. In addition suppose that the following smoothness condition holds true for some diverging real valued sequence $\{r_T\}$,*[12]

$$r_T\Big\|(\nabla Q_T - \nabla Q_\infty)(\hat{\boldsymbol{\theta}}_T) - (\nabla Q_T - \nabla Q_\infty)(\boldsymbol{\theta}_0)\Big\| = o_p(1 + r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|). \tag{1.3}$$

Furthermore, let the second derivative of Q_∞, denoted $\nabla^2 Q_\infty$, satisfy a continuous invertibility condition ensuring, $\Big\|\nabla^2 Q_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})\Big\| \geq c \cdot \|\boldsymbol{\theta}\|$ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$ for some $c > 0$. Finally, let the 'score' satisfy for some real sequence $r_T \to \infty$,

$$\|\nabla Q_T(\boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0)\| = O_p(r_T^{-1}).$$

Then, if $\hat{\boldsymbol{\theta}}_T$ satisfies a Z-estimator formulation $\nabla Q_T(\hat{\boldsymbol{\theta}}_T) = 0$, we obtain the desired result that $r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| = O_p(1)$ as $T \to \infty$. Furthermore, if the normalized 'score' converges in distribution,

$$r_T\Big(\nabla Q_T(\boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0)\Big) \xrightarrow{d} \mathbb{G},$$

then we have that $r_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} -\text{inv}\Big(\nabla Q_\infty\Big(\boldsymbol{\theta}_0, \cdot\Big)\Big)(\mathbb{G}) \quad as\ T \to \infty$.

A modified version of this lemma will play a fundamental role in the theory of SNPII estimation. For now, we retain the idea that it is possible to obtain general results for the convergence rate and asymptotic distribution of smooth extremum estimators in infinite dimensional spaces.

1.3 The Method of Sieves: A Solution to a Statistical Problem

We have argued in the previous section that allowing for an infinite dimensional parameter space might constitute an important step towards generality, yielding

[11] Let us ignore the slight abuse of notation of denoting by 0 the zero element of $\mathbb{R}_\infty$. Also, we could further impose that $\boldsymbol{\theta}_0$ be the unique element of Θ satisfying $\nabla Q_\infty(\boldsymbol{\theta}_0) = 0$. As pointed out by van der Vaart and Wellner (1996), this condition is however unnecessarily restrictive.

[12] Here $\|\cdot\|$ denotes a norm in any given space, lin(Θ) denotes the linear span of Θ and inv(f) denotes the inverse of the operator f.

correct specification axioms less restrictive. However, the simple example in the very beginning of this chapter revealed the difficulties of applying extremum estimator theory to large infinite dimensional spaces. Here we turn to the solution proposed in Grenander (1981).

Written shortly after Grenander's original contribution, Geman and Hwang (1982) derived the first general consistency results for the sieve ML estimator. This paper provided simple yet powerful examples of the failure of classical extremum estimators on large infinite dimensional spaces. Thus revealing the importance of the method of sieves. In their own words,

> *"Techniques for estimating finite dimensional parameters typically fail when applied to infinite dimensional problems. The difficulties encountered [...] are well illustrated by the failure of maximum likelihood in nonparametric density estimation."* in Geman and Hwang (1982)

Let us review the nonparametric density estimation example mentioned above. Let $x_1, ..., x_T$ be an iid sample from an absolutely continuous distribution with unknown pdf denoted $\boldsymbol{\theta}_0(x)$. Then, the ML estimator of $\boldsymbol{\theta}_0$ is designed to maximize $\prod_{t=1}^{T} \boldsymbol{\theta}(x_t)$. When $\boldsymbol{\theta}_0$ is known to belong to small class Θ of probability density functions, then consistency can be obtained under appropriate regularity conditions. However, in the extreme case where nothing is known about $\boldsymbol{\theta}_0$, then Θ will contain the space of discrete pdfs and estimates of $\boldsymbol{\theta}_0$ will consist of discrete density functions with jumps at sample points. Such estimates will not converge to $\boldsymbol{\theta}_0$. Luckily Grenander's method of sieves can be called to offer a solution.

> *"Grenander (1981) suggests the following remedy: perform the optimization with a subset of the parameter space, and then allow this subset to 'grow' with sample size. [...] the resulting estimation is his 'method of sieves'. The method leads easily to consistent nonparametric estimators in even the most general settings, with different sieves giving rise to different estimators."* in Geman and Hwang (1982)

In the context of the above example, the solution offered by the method of sieves consists of the well known *kernel estimator*. The kernel estimator takes values in relatively 'small' subsets of Θ. By letting the *bandwidth* vanish as the sample size increases, the kernel estimator can be shown to converge to $\boldsymbol{\theta}_0$ in a very general space. A similar solution applies to the regression problem introduced in the very beginning of this chapter.

The fundamental idea of the method of sieves is the following. Given a sequence of subsets $\{\Theta_T\}$ called *sieves* of Θ, satisfying $\Theta_T \subseteq \Theta_{T+1} \subseteq \Theta$ for every $T \in \mathbb{N}$, define a *sieve extremum estimator* as follows,

$$\hat{\boldsymbol{\theta}}_T = \arg \min_{\boldsymbol{\theta} \in \Theta_T} Q_T(\boldsymbol{\theta}). \tag{1.4}$$

Now, by letting $\{\Theta_T\}$ 'increase' with sample size and be *dense* in Θ, we can eventually obtain $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$ under appropriate regularity conditions. Note that $\{\Theta_T\}$ is said to be dense in Θ if the closure of its union contains Θ, i.e. $\text{cl}\Big(\cup_{T\in\mathbb{N}}\Theta_T\Big) \supseteq \Theta$. This ensures essentially that for every $\boldsymbol{\theta} \in \Theta$ there exists a sequence $\{\boldsymbol{\theta}_T\}$ of elements $\boldsymbol{\theta}_T \in \Theta_T \, \forall T \in \mathbb{N}$ such that $\boldsymbol{\theta}_T \to \boldsymbol{\theta}$. The following lemma is adapted from White and Wooldrige (1991).

Lemma 1.3.1. *Let $\hat{\boldsymbol{\theta}}_T$ be a sieve extremum estimator as defined in (1.4). Let each sieve Θ_T be compact and satisfy $\Theta_T \subseteq \Theta_{T+1} \subseteq \Theta$, for every T, and let $\{\Theta_T\}$ be dense in Θ. Let, the criterion function Q_T converge uniformly across sieves to Q_∞, i.e. let*

$$\sup_{\boldsymbol{\theta}\in\Theta_T} |Q_T(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0 \;\; as \;\; T \to \infty.$$

Finally, suppose that the limit criterion function Q_∞ is continuous on Θ and that $\boldsymbol{\theta}_0$ is an identifiably unique minimizer of Q_∞. Then, it holds true that $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$ as $T \to \infty$.

Apart the examples considered above, it is not immediately clear why this consistency lemma should allow for more generality than that introduced in Section 1.1 for the consistency of extremum estimators. One condition however, suggests already something new. Notice first that in close resemblance to the extremum estimator theory discussed in Section 1.1, the uniform convergence of $\{Q_T\}$ to Q_∞ also plays an important role in the consistency of the sieve estimator. This time however, the uniform convergence occurs *'across sieves'*. This apparently innocent statement, carries important implications concerning the 'size' of Θ.

Recall from Section 1.1 that there are two important ways of deriving the uniform convergence of $\{Q_T\}$ to Q_∞. One related to *generic uniform convergence* theorems and another related to *empirical process theory*. In general however, both require Θ to satisfy some form of total boundedness or finite complexity.[13]

Remark 1.3.1. *In infinite dimensional spaces, assumptions of total boundedness or finite complexity are extremely restrictive. As a result, the standard theory of uniform convergence (which relies on such assumptions) can not be used to obtain a meaningful general theory of extremum estimation on infinite dimensional spaces.*

Fortunately, both the theory of *generic uniform convergence* and *empirical processes* can be adapted to hold on unbounded spaces of infinite complexity as long as uniform convergence is required to hold only across finite dimensional subsets of Θ. This is precisely the form of convergence required for the consistency of sieve extremum estimators in Lemma 1.3.1 above.

[13] Complexity is measured in terms of the *covering number* or *entropy* of the set Θ.

We end this section with a brief account of the existing results on consistency, convergence rate and asymptotic distribution of sieve extremum estimators. The main message is that while the consistency theory of sieve extremum estimators is fairly well established and complete, the theory of convergence rates and weak convergence is much less developed. In essence, there are no theorems that derive convergence rates or weak convergence of general sieve extremum estimators.

Important contributions for the consistency theory of sieve extremum estimators and semi-nonparametric models include Geman and Hwang (1982) for the sieve ML estimator, Gallant and Nychka (1987) and Gallant (1987) for M-estimators of semi-nonparametric models (requiring compactness of Θ), and White and Wooldrige (1991) for the general sieve extremum estimator. See Chen (2007) for further references and a consistency theorem for sieve extremum estimators.

Some first important results on the convergence rate of the sieve ML estimator were obtained by Wong and Severini (1991) for compact spaces. Subsequent developments included Birgé and Massart (1993) and Shen and Wong (1994) that provided the first results on the convergence rate of sieve M-estimators for iid data. Relevant literature on the convergence rates of the sieve M-estimator includes also Van de Geer (1995) and Birge and Massart (1998) and Chen and Shen (1998). Results on the sieve ML estimator are also available in Van de Geer (1993) and Wong and Shen (1995).[14]

Asymptotic normality results for sieve estimators are still scarce and in general apply only to either series least squares estimators or to the finite dimensional parametric part of semi-parametric models; see e.g. Andrews (1991), Gallant and Souza (1991), Newey (1994, 1997), Zhou et al. (1998) and Huang (2003) for results on series least-squares estimators, and Wong and Severini (1991), Gallant and Souza (1991), Shen (1997), Chen and Shen (1998) and Chen et al. (2003) for both two-step and simultaneous M-estimators. It is also important to point out that these results have been generally obtained under quite restrictive conditions on the heterogeneity and dependence of the data. Once more, the reader is referred to Chen (2007) for further details.

Finally, we offer a very brief remark that aims to clarify the relation between the method of sieves and Semi-Nonparametric (SNP) models.

Method of Sieves and SNP Models

SNP models were introduced in the econometrics literature by Gallant (1981). An SNP model is not only a collection of probability distributions $\mathcal{D}_\Theta := \{\mathrm{D}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$, but also, a collection of well-defined sub-models $\mathcal{D}_{\Theta_T} := \{\mathrm{D}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_T\}$ satisfying $\mathcal{D}_{\Theta_T} \subseteq \mathcal{D}_\Theta$ which define a restriction on $\mathcal{D}_\Theta$ for every $T \in \mathbb{N}$.

[14] Other results exists also for specific sieves and criterion functions; see Chen (2007).

In the method of sieves, the parameter space Θ is unbounded and infinite dimensional, but the estimator $\hat{\boldsymbol{\theta}}_T$ is restricted to take values in a subset Θ_T. More generally, the sequence of estimators $\{\hat{\boldsymbol{\theta}}_T\}$ takes values on finite dimensional compact subsets $\{\Theta_T\}$. Clearly, the exact same restriction is also imposed in the SNP literature. There the model of interest is the large collection $\mathcal{D}_\Theta$ of probability distributions indexed by $\boldsymbol{\theta} \in \Theta$, but estimation takes place using 'smaller' models $\mathcal{D}_{\Theta_T}$. Following Chen (2007) we shall often refer to *sieve estimators of semi-nonparametric models*.

Remark 1.3.2. *The design of semi-nonparametric models is quite flexible. Most common is the formulation of sieves Θ_T spanned by a basis vector of increasing dimension. For example, Gallant (1981) used a truncated Fourier series (with truncation order diverging with T) to approximate very general functions. The same idea could proceed with various polynomials of increasing order, splines, neural networks, and others.*

Section 1.4 below clarifies these ideas and reviews relevant approximation methods. The very instructive work of Judd (1992, 1998) and the excellent review of Chen (2007) provide many more details.

1.4 Approximation Theory

A fundamental characteristic of the theory of sieve estimation reviewed above concerns the denseness of the sieves $\{\Theta_T\}$ on the parameter space of interest Θ. This requires that every element $\boldsymbol{\theta} \in \Theta$ be arbitrarily well approximated by sequences $\{\boldsymbol{\theta}_T\}$ in the sieves $\{\Theta_T\}$. In essence, this is a problem of approximation in Θ. *Approximation Theory* is thus a fundamental component of the theory of sieve estimation and SNP models. In what follows the interested reader can find a brief review of this literature.

Our journey begins with Mãdhava of Sañgamãgrama (1350–1425), one of the greatest mathematicians of the middle ages. Usually regarded as the founder of the *Kerala School of Astronomy and Mathematics*, he was responsible for revolutionary work with infinite series. Indeed, the first Taylor series expansions of several trigonometric functions are attributed to him. While important work on series expansions and rational approximations continued in the Kerala School for a long time after his death, it was only two centuries later that the Scottish mathematician James Gregory published several Maclaurin series in his work *"Vera Circuli et Hyperbolae Quadratura"* in 1667. Some regard Gregory as the "inventor" of Taylor series.[15]

[15] Apparently, James Gregory wrote to John Collins, secretary of the Royal Society, on February 15, 1671, to tell him of the result. The first draft of Gregory's discovery is preserved on the back

A general method of obtaining approximating series came only five decades later. This result arrived in the year of 1715 in Brook Taylor's *"Methodus Incrementorum Directa et Inversa"* in the form of some formulas that are now known as the much celebrated *Taylor's Theorem*. Curiously enough, this result would remain largely unknown until found by the mathematician and astronomer Joseph-Louis Lagrange in 1772. The recognition of its importance was made clear in his statement that called Taylor's Theorem *"the main foundation of differential calculus"*. It is probably due to such initial obscurity that Colin Maclaurin published soon after his Maclaurin series expansions, which turned out to be only special cases of those of Taylor.[16]

Remark 1.4.1. *Consider the space $\mathbb{C}^\infty(\mathcal{X})$ of all real-valued functions that are infinitely differentiable on the open interval $\mathcal{X} \subset \mathbb{R}$. A subset of so-called "analytic functions", denoted $\mathbb{C}^\omega(\mathcal{X}) \subset \mathbb{C}^\infty(\mathcal{X})$, can be represented as an infinite power series. In other words, if $f : \mathbb{R} \to \mathbb{R}$ is analytic on $\mathcal{X}$, then, for every $x_0 \in \mathcal{X}$, it holds true that $f(x) = \sum_{k=0}^{\infty} \theta_k (x - x_0)^k$ for every x in a neighborhood of x_0. The space $\mathbb{C}^\omega(\mathcal{X})$ is thus spanned by the infinite sequence of power monomials $\{1, x, x^2, ...\}$. Taylor's Theorem showed that the power series representation of $f \in \mathbb{C}^\omega(\mathcal{X})$ holds with $\theta_k = f^{(k)}(x_0)/k!$ where $f^{(k)}(x_0)$ denotes the k^{th} derivative of f at x_0. In the context of function approximation with a truncated power series $p_K(x) = \sum_{k=0}^{K} \theta_k (x - x_0)^k$, setting $\theta_k = f^{(k)}(x_0)/k!$ defines the unique polynomial that 'matches' the first K derivatives, i.e. $p_K(x_0) = f(x_0)\,,\ p'_K(x_0) = f'(x_0)\,,\ \ldots\,,\ p_K^{(K)}(x_0) = f^{(K)}(x_0)$. Taylor's coefficients thus provide optimal approximations w.r.t. the semi-norm $\rho(f - p_K) = \sum_{k=0}^{K} |f^{(k)}(x_0) - p_K^{(k)}(x_0)|$, i.e. they minimize $\rho(f - p_K)$. In essence, given $f \in \mathbb{C}^K(\mathcal{X})$, the polynomial $p_K \in \mathbb{P}_K(\mathcal{X})$ is the unique best approximation (w.r.t. ρ) to f from $\mathbb{P}_K(\mathcal{X})$.*

Almost a century later, Augustin-Louis Cauchy and Joseph-Louis Lagrange derived explicit formulas for the remainder of function approximation by truncated power series with Taylor coefficients. These formulas gave truncated power series a further "approximation flavor" and became known as the Cauchy and the Lagrange remainders respectively.

Still, by describing a truncated Taylor series, as the polynomial of order K that minimizes a certain distance, the natural question to be asked is *which polynomials minimize other distances of interest*? In 1779, Edward Waring discovered a method to find the (unique) K-th order polynomial that interpolates a function at $K + 1$

of a letter he received on 30 January, 1671, from an Edinburgh bookseller.

[16] Nonetheless, the achievements of Colin Maclaurin since a very young age earned him the admiration of several contemporary mathematicians. In 1725, the great Sir Isaac Newton actually offered to pay a salary to Colin Maclaurin, from his own budget, in a letter addressed to John Campbell as means of persuading him to accept Colin Maclaurin for an appointment at the University of Edinburgh.

distinct points. This polynomial, that was later rediscovered independently by the the Swiss mathematician and physicist Leonhard Euler in 1783, came to be known as *Lagrange polynomial*.

Remark 1.4.2. *Given a function f, the Lagrange polynomial $p_K = \sum_{k=0}^{K} \theta_k x^k \in \mathbb{P}_K$ is the unique polynomial of order K that minimizes the 'interpolation seminorm' $\rho(f - p_K) = \sum_{k=0}^{K} |f(x_k) - p_K(x_K)|$, i.e. the unique polynomial that satisfies, $f(x_0) = p_K(x_0)$, ..., $f(x_K) = p_K(x_K)$. In this context, $\{x_k\}_{k=0}^{K}$ are known as collocation nodes and p_K as an interpolating polynomial of f. Waring found that the interpolating polynomial is given by $p_K(x) = \sum_{k=0}^{K}(f(x_k)/A_k(x_k))A_k(x)$ where $A_k(x) = \prod_{j=0,j\neq k}^{K}(x - x_j)$ and $A_k(x_k) = \prod_{j=0,j\neq k}^{K}(x_k - x_j)$.*

Just two years after Euler's work on interpolating polynomials, the French mathematician Adrien-Marie Legendre enriched the possibilities of function approximation with his *"Recherches sur l'attraction des sphéroïdes homogènes"* published in 1785. His work opened the door to the approximation of functions using linear combinations of orthogonal polynomials. Although Legendre's interest lied on providing solutions to differential equations, his polynomials turned out to have interesting approximation properties.

Remark 1.4.3. *Legendre functions are solutions to the Legendre's differential equation $(d/dx)[(1 - x^2)(d/dx)P_n(x)] + n(n + 1)P_n(x) = 0$. The solutions for n integer with $P_n(1) = 1$ form the sequence of Legendre polynomials. These polynomials can be obtained according to the recurrence relation $P_0(x) = 1$, $P_1(x) = x$ and $(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$, and are orthogonal on $[-1, 1]$ w.r.t. the weighting function $w(x) = 1$, i.e. $\int_{-1}^{1} P_n(x)P_m(x)dx = 0 \,\forall n \neq m$.*

Few approximation methods can however challenge the revolutionary importance of the discovery that was to take place, just two decades later, by the hand of Joseph Fourier in his *"Mémoire sur la propagation de la chaleur dans les corps solides"*, published in 1807. This work focused on finding a general solution to a partial differential equation known as the *heat equation*. His solution took the form of a series of trigonometric functions. Fourier's series turned out to have enduring influence in many areas of science. While the famous Leonhard Euler and Daniel Bernoulli had previously investigated the properties of such series, it was Fourier that claimed the vastness of its application in terms of approximating large classes of functions. Fourier's work was nonetheless received with some criticism. When submitted to a competition, a board of examiners which included his own professor Joseph Lagrange, as well as, Pierre-Simon Laplace and Adrien-Marie Legendre, stated about Fourier's result that *"The manner in which the author arrives at these equations is not exempt of difficulties and [...] his analysis to integrate them still*

leaves something to be desired on the score of generality and even rigor".[17]

Remark 1.4.4. *When it exists, a Fourier series of a function f on $[-\pi, \pi]$ takes the form $a_0/2 + \sum_{k=1}^{\infty}[a_k \cos(kx) + b_k \sin(kx)]$ with $a_k = 1/\pi \int_{-\pi}^{\pi} f(x)\cos(kx)dx$ and $b_n = 1/\pi \int_{-\pi}^{\pi} f(x)\cos(kx)dx$. The Riesz–Fischer theorem, proved independently by Ernst Fischer and Frigyes Riesz in 1907, provided a definite representation result for the class of L_2 functions in terms of Fourier series. Truncated Fourier series are also useful in approximating important classes of periodic and non-periodic functions, including functions with certain discontinuities. Given a truncated series $s_K(x) = a_0/2 + \sum_{k=1}^{K} a_k \cos(kx) + b_k \sin(kx)$, Fourier's coefficients are optimal w.r.t. the L_2-norm, in the sense that they minimize $\left[\int_{-\pi}^{\pi}[f(x) - q(x)]^2 dx\right]^{1/2}$.*

By the mid 19th century, a very important method of function approximation which also admits a formulation in terms of a series of trigonometric functions would be introduced by the great Russian mathematician Pafnuty Chebyshev in his *"Théorie des mécanismes connus sous le nom de parallélogrammes"* in 1854.[18] Over time, approximation of functions by linear combinations of Chebyshev polynomials became extremely famous due to their important optimality properties in several applications. Just like Legendre polynomials, Chebyshev polynomials benefited from the properties of orthogonality and the ability to be obtained in a simple recursive fashion. A further advantage exclusive to Chebyshev polynomials is however that their roots (when used as nodes in polynomial interpolation) turn out to minimize *Runge's phenomenon*, documented five decades later in Carl Runge's *"Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten"* in 1901. Runge showed that, in some applications, the increase of approximation order in polynomial interpolation might actually decrease accuracy. This is due to increased oscillation in the polynomial approximation. This oscillation can however be minimized by using the roots of Chebyshev polynomials as collocation nodes.

Remark 1.4.5. *Chebyshev polynomials are defined on $[-1, 1]$ and obtained using the recursion formula, $T_0(x) = 1$, $T_1(x) = x$ and $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ for $n > 1$. These polynomials are orthogonal w.r.t. the weight function $w(x) = (1 - x^2)^{-1/2}$, i.e. $\int_{-1}^{1}(1 - x^2)^{-1/2}T_m(x)T_n(x)dx = 0$ for every $n \neq m$.*[19] *Chebyshev polynomials form a complete orthogonal basis of a Sobolev space and are related to*

[17] Apparently, Fourier thought that virtually all functions could be approximated by a Fourier series. This is however false. Andrey Kolmogorov's "Une série de Fourier-Lebesgue divergente presque partout" published 1922, provides a well known counter example of a Lebesgue-integrable function whose Fourier series diverges almost everywhere.

[18] The "Théorie des mécanismes connus sous le nom de parallélogrammes" is one of the many works that Chebyshev wrote in French.

[19] In Legendre polynomials, the constant weight function implies that errors occurring close to the borders of $[-1, 1]$ are actually given less weight (only one-sided errors are present) than errors

Fourier cosine series by a change of variable. Hence, results derived for the latter apply appropriately to the former. Given a function $f \in \mathbb{C}^n[-1,1]$, the K-th order Chebyshev approximation converges uniformly at rate $O(\ln(K)K^{-n})$ to f.

In the two decades following Chebyshev's *Théorie des mécanismes*, two new approximation polynomials were introduced that would remain equally popular until present times. The first was introduced by Charles Hermite's *"Sur un nouveau développement en série de fonctions"* in 1864.[20] The second by Edmond Nicolas Laguerre in *"Sur l'intégrale* $\int_x^{+\infty} x^{-1}e^{-x}dx$*"* published in 1879. These polynomials are known in present times as Hermite and Laguerre polynomials respectively.

Remark 1.4.6. *Both Hermite and Laguerre polynomials are obtained according to* $H_K(x) = (-1)^K \exp(x^2)\frac{\partial^K}{\partial x^K}\exp(-x^2)$ *and* $L_K(x) = \exp(x)/n!\frac{\partial^K}{\partial x^K}(x^K\exp(-x))$ *respectively. These polynomials are orthogonal w.r.t. the weighting functions* $w(x) = \exp(-x^2)$ *and* $w(x) = \exp(-x)$ *respectively. Due to their weighting functions, these polynomials are especially suited to approximate functions on* $\mathbb{R}$ *and* $\mathbb{R}_0^+$ *respectively. Hermite and Laguerre polynomials play an important role in Gaussian quadrature methods involving the approximation of integrals of functions that decay exponentially.*

In 1892 the theory of approximation by rational polynomials was introduced in Henri Padé's "Sur la répresentation approchée d'une fonction par des fractions rationelles".

Remark 1.4.7. *A Padé approximant* $r_{m,n}$ *of a function* f *at a point* x_0 *takes the form,* $r_{m,n}(x) = \frac{p_m(x)}{q_n(x)} = \frac{\sum_{i=0}^{m}\theta_i x^i}{1+\sum_{j=1}^{n}\beta_j x^j}$*, where the* θ_i*'s and* β_j*'s are derived from the condition* $p^{(k)}(x_0) = (fq)^{(k)}(x_0)$ *for* $k = 0, ..., m+n$*. Similarly to a Taylor series, the Padé approximant is the rational polynomial that minimizes the semi-norm* $\rho(f - r_{m,n}) = \sum_{k=0}^{K}|f^{(k)}(x_0) - r_{m,n}^{(k)}(x_0)|$ *so that a Padé approximant* $r_{m,n}$ *also agrees with* f *and its derivatives at* x_0*. Padé approximants are not only* ρ*-optimal, they often converge where Taylor series do not (e.g. close to poles and other singularities).*

The 19th century history of function approximation was a rich one and it would not come to an end without the introduction of the much celebrated and highly influential approximation theorem of the German mathematician Karl Weierstrass in *"Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen"* in 1885. As mentioned above, uniform convergence of polynomials to smooth functions in $\mathbb{C}^n[-1,1]$ can be obtained using e.g. Chebyshev polynomials. An important question thus remained: wether the larger set of continuous

occurring close to the center. Chebyshev's weighting function counteracts this effect. Legendre polynomials offer generally a poorer approximation than Chebyshev's.

[20] Laplace and Chebyshev had already studied the properties of Hermite polynomials sometime earlier.

functions $\mathbb{C}[a,b]$ contains pathological functions for which such uniform convergence does not hold. The famous theorem of Karl Weierstrass proved essentially that any continuous real valued function defined on an interval $[a,b]$ can be arbitrarily well approximated in sup norm by a polynomial function. The generality of Weierstrass's theorem was far reaching and profound, but it was only in the first half of the 20th century that today's well known (and much more general) version of the theorem arrived by the hand of the American mathematician Marshall Stone in his *"Applications of the Theory of Boolean Rings to General Topology"* and *"The Generalized Weierstrass Approximation Theorem"* in 1937 and 1948 respectively. Due to Stone's work, present formulations of the *Stone-Weierstrass Theorem* hold for functions defined on general compact Hausdorff spaces.[21]

Remark 1.4.8. *Weierstrass's Theorem was given a constructive proof by the Russian mathematician Sergei Bernstein in 1912. In particular, Bernstein showed that given a function $f \in \mathbb{C}[0,1]$, the polynomial $B_n(x) = \sum_{k=0}^{n} \theta_k x^k (1-x)^{n-k}$ with $\theta_k = f(k/n)\frac{n!}{k!(n-k)!}$ for $k = 0, ..., n$, converges uniformly to f as $n \to \infty$. Bersntein polynomials have the further important property that the derivatives of B_n converge uniformly to the derivatives of f.*

On a more theoretical note, the early 20th century was also witness to other great developments in the world of approximation theory. Two such developments of great importance were the study of *Schauder basis* and of spaces with the *approximation property*. Schauder basis extended the usual notion of Hamel basis (named after Georg Hamel, a doctoral student of David Hilbert) from finite to infinite-dimensional spaces. While spaces equipped with a Hamel basis describe its elements as a linear combination of finitely many basis vectors, Schauder basis allow vectors to be obtained as linear combinations of infinitely many elements of the basis. Schauder basis had been studied earlier in 1909 by Alfréd Haar (also a student of David Hilbert) in his work on the *Haar basis* in *"Zur Theorie der orthogonalen Funktionensysteme"*. However, Schauder basis are named after Juliusz Schauder for his work *"Zur Theorie stetiger Abbildungen in Funktionalraumen"* in 1927 and *"Eine Eigenschaft des Haarschen Orthogonalsystems"* in 1928.

Remark 1.4.9. *The theory of Schauder spaces is extremely relevant to the understanding of which infinite dimensional spaces can be well approximated by a sequence of smaller spaces obtained as the linear span of an increasing number of basis vectors. Schauder basis plays an important role in the theory contained in this thesis.*

A famous problem posed by the Polish mathematician Stefan Banach asked whether every separable Banach space had a Schauder basis. In a paper published

[21] It it thus intuitively clear that the space of continuous functions (with sup norm) defined on a compact Hausdorff space is separable (i.e. it contains a dense countable subset).

in 1973, Per Enflo stunned the world by providing a first negative answer to Banach's question in the form of a counter example. Enflo's example solved also the closely related *Mazur's Goose problem* and the *Approximation problem* of Alexander Grothendieck.[22]

With the advent of the computer and its increasing power, the second half of the 20th century witnessed a rapid development of computationally intensive methods for approximating functions. Some of these have come to shape quite substantially both theoretical and applied work in several areas of science and engineering. Essentially, the growing computing power has made it practical to turn a single approximation problem on a domain $\mathcal{X}$, into several 'smaller' approximation problems on partitions of $\mathcal{X}$. The partitioning of the original domain defines a so-called *mesh*. Approximation is then shown to improve as the size of the elements of the mesh becomes smaller. In its simpler form, approximation takes place using piecewise linear functions by collocation methods.[23] Under certain regularity conditions, a more promising approach uses higher order polynomials on each element of the mesh and ensures 'smoothness at transition points'. In statistics, this method is essentially known as the *method of smoothing splines*, which originated in the contributions of Whittacker (1923), Schoenberg (1964) and Reinsch (1967). See de Boor (1978), Schumaker (1981) and Powell (1981) for reviews of the *spline approximation* and *smoothing splines*.

Remark 1.4.10. *Spline is the name generally given to a function that takes the form of a piecewise polynomial of degree (at most) K in a domain $\mathcal{X} \subseteq \mathbb{R}$ and ensures the continuity of its $K-1$ derivative. A spline function of degree K on the mesh $[\xi_{i-1}, \xi_i]$, $i = 1, ..., N$, can be shown to admit the general form $s_K(x) = \sum_{k=0}^{K} \beta_k x^k + (1/K!) \sum_{k=1}^{N-1} \rho_k (\max[0, x - \xi_k])^K$. In general, the best spline approximation s_K to a function $f \in \mathbb{C}^{K+1}[a, b]$ satisfies $\|f - s\|_\infty = O(h^{K+1})$ where h denotes the size of the largest mesh element $h := \max_i |\xi_{i-1} - \xi_i|$. Depending on the smoothness of f, a higher or lower order spline might be desired (see Powell (1981) for this and many other results). Finally, the 'smoothing spline' typically obtains the coefficients $\{\beta_k\}$ and $\{\rho_k\}$ by minimizing a least squares criterion function with a smoothness*

[22] The 'Goose problem' was stated by Stanislaw Mazur as the problem number 153 of the famous *Scottish book*. This book was used to state unsolved problems by the group of famous mathematicians, of the Polish Lwów School of mathematics, that met regularly in the *Scottish Cafe* of Lwów. This group included Stefan Banach, Kazimierz Kuratowski, Stanisław Mazur, Juliusz Schauder and Stanislaw Ulam. With each problem came a prize offered to the first person to solve it. The famous group meetings ended with the German invasion of Poland. For solving Mazur's problem, Enflo was offered in 1972 a live goose, the prize promised by Mazur in 1936.

[23] This is essentially the idea of the linear *finite element method* (FEM). The FEM originated in the work of Alexander Hrennikoff and Richard Courant in 1942 and 1942 respectively. Higher-order polynomials and minimization of *Galerkin weights* is most common in the FEM literature.

penalty, $N^{-1}\sum_{i=1}^{N}\left(f(x_i)-s_K(x_i)\right)^2+\lambda\int\left(s_K^{(K)}(x_i)\right)^2dx$*. Denseness and convergence theorems for general splines apply naturally to smoothing splines as well.*

We close this section with the recent development of *artificial neural networks*. Originated in the cognitive science literature, artificial neural networks provide a very flexible function approximation method that has gained popularity as a semi-nonparametric modeling tool. In the strict perspective of approximation theory, important developments existed already since Hecht-Nielsen (1987) which used Kolmogorov's superposition theorem to show that *single hidden-layer feedforward artificial neural networks* could be used to approximate arbitrary continuous mappings. Some initial important statistical foundations were laid down in White (1989a,b) (for the special case of single hidden-layer feedforward network models) which documented that *''the excitement evident across such disciplines as psychology, computer science, linguistics, and engineering is founded on the demonstrated success in solving a diversity of difficult problems that had previously withstood conventional attacks"*. Further developments on the approximation properties of artificial neural networks (ANN) arrived immediately after in Hornik et al. (1989) and White (1990) which showed not only that ANNs can approximate arbitrary Borel measurable maps, but also, that the approximation is, in the field's language, *learnable*. Essentially, this established the use of ANNs as a valid and very general statistical procedure for estimation of functions. Hornik et al. (1989), Gallant and White (1992) and Hornik et al. (1994) showed that *learnable* ANNs with smooth *squashing functions* approximate not only arbitrary smooth functions, but their derivatives as well; see also, Chen and White (1998) and Chen and White (1999). In time-series analysis, the importance of ANNs was reenforced by the availability of several results on the geometric ergodicity and stationarity of autoregressive ANN models (see e.g. Trapletti et al. (1998)).

Remark 1.4.11. *A single hidden layer feedforward artificial neural network takes the general form* $\boldsymbol{\beta}_0+\sum_{k=1}^{K}\boldsymbol{\beta}_k\phi(\rho_k+\gamma_k x)$ *where* ϕ *is called a 'squashing function' and typically takes the form of a sigmoid function, a cumulative distribution function or a logistic. Linear and quadratic components can also be included. The ANN framework is hence quite general and takes as special cases many of the previously mentioned approximation methods, e.g. Fourier series for sine and cosine squashing functions. Clearly, the parameters* β_k*,* ρ_k *and* γ_k *are not identified. Estimation can nonetheless proceed in a sequential or 'online learning' way (see White 1989b). Typically this is done by minimizing some form of weighted least squares criterion function. Gallant and White (1988a) showed that the sigmoid ANN with cosine squashing is dense (in* sup *norm) in the space of continuous functions defined on a compact subset* $\mathcal{X}\subset\mathbb{R}^d$*. Hornik et al. (1994) generalized this result for any sigmoid*

activation function. Results on convergence rates include Hornik et al. (1994) and Chen and White (1999) for approximation of functions on Sobolev and L^2 spaces. See Judd (1998) and Chen (2007) for a review of other results.

1.5 Indirect Inference: Learning from Auxiliary Statistics

The method of sieves introduced in Section 1.3 allowed us to deal with infinite dimensional spaces of unbounded complexity. There, we reviewed results that suggest the possibility (at least in theory) of conducting statistical inference on such large parameter spaces. However, we have not made any comments on the practical implementation of such procedures. We now turn our attention to this issue.

Recall that the consistency of the sieve extremum estimator discussed in Section 1.3 relied fundamentally on the increasing complexity of the sieves. Implicitly, the assumption was also made that an estimator taking values on such sets is available.

Remark 1.5.1. *If the method of sieves is to be of any use, an estimator must be available that is practical to work with. The availability of a 'practical' sieve estimator might not be a matter of concern in the simple regression and density estimation cases considered until now. Outside these simpler cases however, complications are likely to occur.*

Consider the nonlinear cross-sectional regression problem introduced in the beginning of this chapter. In principle, it is not difficult to work with sieve estimators of the type $\hat{\boldsymbol{\theta}}_T(x) = \sum_{k=0}^{K_T} \beta_k x^k$ where $K_T \to \infty$ as $T \to \infty$. Indeed, such a sieve estimator takes values in sieves Θ_T that are spanned by the basis vectors $\Theta_T \subseteq \text{lin}\{1, x, x^2, ..., x^{K_T}\}$ for every T. Furthermore, given the results of Section 1.3, we know that (under appropriate regularity conditions) an estimator designed in this way can be consistent to a parameter $\boldsymbol{\theta}_0$ lying on a space Θ of continuous functions in x. Until here everything seems to work well. However, consider now the nonlinear dynamic model introduced in the beginning of this chapter,

$$\mathbf{x}_t = \boldsymbol{\theta}_0(\mathbf{x}_{t-1}) + \boldsymbol{\epsilon}_t$$

where ϵ_t is a vector of innovations and the vectors $\mathbf{x}_t$ contain both observed and latent variables. In such a setting, difficulties can be expected when applying the sieve estimation methodology to estimate $\boldsymbol{\theta}_0$.

Remark 1.5.2. *Even in relatively simple dynamic models, classical estimators such as maximum likelihood and method of moment estimators might be hard to derive. If this is true for relatively simple models, not much can be expected from dynamic models whose complexity must increase with T.*

Below, we review a solution to our problem that goes by the name of *indirect inference* (II). This solution is available for finite dimensional parameter spaces only. Hence, for the time being, we leave the sieves method aside.

With the availability of increased computational power, the 1980's witnessed a growing interest in simulation-based estimators. On finite dimensional parameter spaces, such estimators offer an alternative to classical estimators and are especially appealing when (due to model complexity or others) the latter fail to be available. This literature includes simulation-based extensions of classical estimators such as *simulated maximum likelihood*, *simulated method of moments* and others; see Gourieroux and Monfort (1996), Dave and Dejong (2007) and Ruge-Murcia (2007) for reviews of this literature. The II principle underlying these techniques was introduced in Gourieroux et al. (1993); see also Smith (1993).

As we shall see, the principle of II does more than just describing the fundamental ideas behind simulation-based estimators. It provides a general setting for statistical inference that relies on *auxiliary statistics* or *auxiliary estimators* (regardless of a possible need for simulations, or not). In essence, it deals with estimators that are defined as functionals of other estimators.

Following Gourieroux et al. (1993), let $\mathbf{x_T} := (\mathbf{x}_1, ..., \mathbf{x}_T)$ denote a T-period sample of observed data. Furthermore, suppose that the distribution of $\mathbf{x_T}$ is implicitly defined by the following dynamic model,

$$\begin{aligned} \mathbf{x}_t &= \mathbf{h}\Big(\mathbf{x}_{t-1}\,,\, \mathbf{z}_t\,,\, \boldsymbol{\theta}_0\Big) \\ \mathbf{z}_t &= \mathbf{g}\Big(\mathbf{z}_{t-1}\,,\, \boldsymbol{\epsilon}_t\,,\, \boldsymbol{\theta}_0\Big) \quad , \qquad t \in \mathbb{Z}, \end{aligned}$$

where $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^p$, $\mathbf{z}_t$ denotes a vector of latent variables, and $\boldsymbol{\epsilon}_t$ a vector of innovations with known distribution $\mathrm{D}_{\boldsymbol{\epsilon}}$. Suppose that we are interested in conducting inference on $\boldsymbol{\theta}_0$, but that classical estimators are not available. Then, if all the features of the dynamic model are known (except for $\boldsymbol{\theta}_0$) we can still proceed to estimate $\boldsymbol{\theta}_0$ by appealing to the principle of II. In particular, by 'drawing' from $\mathrm{D}_{\boldsymbol{\epsilon}}$, we can obtain sequences $\tilde{\boldsymbol{\epsilon}}_1, ..., \tilde{\boldsymbol{\epsilon}}_T$ and use these to simulate sequences of 'artificial data', denoted $\tilde{\mathbf{x}}_{\mathbf{T}}(\boldsymbol{\theta})$, according to,

$$\begin{aligned} \tilde{\mathbf{x}}_t &= \mathbf{h}\Big(\tilde{\mathbf{x}}_{t-1}\,,\, \tilde{\mathbf{z}}_t\,,\, \boldsymbol{\theta}\Big) \\ \tilde{\mathbf{z}}_t &= \mathbf{g}\Big(\tilde{\mathbf{z}}_{t-1}\,,\, \tilde{\boldsymbol{\epsilon}}_t\,,\, \boldsymbol{\theta}\Big) \quad , \qquad t \in \mathbb{N}, \end{aligned}$$

for any $\boldsymbol{\theta} \in \Theta$. By repeating this procedure, we can obtain multiple simulated sequences $\tilde{\mathbf{x}}^1_{\mathbf{T}}(\boldsymbol{\theta}), ..., \tilde{\mathbf{x}}^S_{\mathbf{T}}(\boldsymbol{\theta})$. Now, the idea of II is to make use of *auxiliary estimators* $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})$ to 'describe' the properties of both observed data $\mathbf{x_T}$ and simulated data $\tilde{\mathbf{x}}^s_{\mathbf{T}}$, and then, to 'search' for the parameter $\boldsymbol{\theta} \in \Theta$ that makes simulated data $\tilde{\mathbf{x}}^s_{\mathbf{T}}(\boldsymbol{\theta})$ as 'similar' as possible to observed data $\mathbf{x_T}$ (as judged by the auxiliary statistics $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,s}$).

For concreteness, let $\hat{\boldsymbol{\beta}}_T$ denote an estimator in $\mathbb{R}^q$ that 'describes' observed data $\mathbf{x_T}$, and $\tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})$ denote the corresponding estimator obtained from the s^{th} sequence of simulated data $\tilde{\mathbf{x}}^s_{\mathbf{T}}(\boldsymbol{\theta})$. For example, $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})$ might consist of sample moments, or correspond to estimators of a simpler model describing the dynamic properties of the data. All that matters is that they provide a 'rich enough' characterization of the distributions of $\mathbf{x_T}$ and $\tilde{\mathbf{x}}^s_{\mathbf{T}}$. In essence this means that $\tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})$ should converge in an appropriate fashion to a singleton limit $\boldsymbol{\beta}^*(\boldsymbol{\theta})$ that satisfies $\boldsymbol{\beta}^*(\boldsymbol{\theta}) \neq \boldsymbol{\beta}^*(\boldsymbol{\theta}')$ for every $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ in Θ.

As a deterministic function of $\boldsymbol{\theta}$, the limit $\boldsymbol{\beta}^*$ is called the *binding function*. The binding function plays an essential role in II estimation as its properties determine the ability to conduct inference on Θ through the use of auxiliary statistics.

Finally, we define the II estimator $\hat{\boldsymbol{\theta}}_T$ as,

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} \; \mu\Big(\hat{\boldsymbol{\beta}}_T \;,\; 1/S \sum_{s=1}^{S} \tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})\Big),$$

where μ is some 'divergence' that measures some notion of 'distance' between $\hat{\boldsymbol{\beta}}_T$ and $1/S\sum_{s=1}^{S}\tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})$. For the special case of a 'quadratic weighted divergence' Gourieroux et al. (1993) show that, under appropriate regularity conditions, $\hat{\boldsymbol{\theta}}_T$ converges to $\boldsymbol{\theta}_0$. The same authors show also that $\hat{\boldsymbol{\theta}}_T$ is $\sqrt{T}$ consistent and asymptotically normal; see also Smith (1993).

We finish this section with a couple of notes on the generality of the II procedure that are important for the theory that follows. First, note that there is no need for auxiliary estimators to be parametric. In fact, non-parametric auxiliary estimators might be preferable in several occasions; see e.g. Billio and Monfort (2003) and Nickl and Pötscher (2009). Second, depending on the 'objective' of the econometric exercise, the requirement of correct specification can be weakened or even eliminated. Indeed, a considerable body of literature has been devoted to the study of *(i)* the properties of II estimators in misspecified models, including the properties of its *'indirect pseudo-true limit'* $\boldsymbol{\theta}_0^*$ and the role of II estimators in testing encompassing hypothesis (Dhaene et al. 1998), *(ii)* the development of robust II estimators (Genton and Ronchetti 2003) and *(iii)* the use of II estimators in semi-parametric models (Dridi and Renault 2000).

The theory in this thesis differs from the above mentioned literature in the following aspects. First, it allows not only the auxiliary parameter space to be infinite dimensional (as in Billio and Monfort (2003) and Nickl and Pötscher (2009)) but also the parameter space of interest Θ to be infinite dimensional. Second, interest lies not on a parametric subset of $\boldsymbol{\theta}_0$ (as in Dridi and Renault 2000) but on the 'unpartitioned' parameter $\boldsymbol{\theta}_0$. The parameter $\boldsymbol{\theta}_0$ of interest defines completely the 'true' distribution. Third, interest lies not in potential effects of misspecification

(as in Genton and Ronchetti 2003) but on reducing the restrictive nature of correct specification axioms by allowing Θ to be very large.

As explained in Section 1.2 however, if the objective is that of conducting inference on a 'true' parameter $\boldsymbol{\theta}_0 \in \Theta$ then, an axiom of correct specification must forcefully hold. If we are to entertain the idea that the assumption of correct specification has any reasonable possibility of holding true, then Θ must be allowed to be infinite dimensional and of infinite complexity. In the next section, we shall finally introduce a technique that promises to deal well with these requirements.

1.6 Semi-Nonparametric Indirect Inference: Econometric Motivation

The preceding sections have suggested a way of dealing simultaneously with the problems of *excessive simplicity* and *excessive complexity* alluded to in the very beginning of this chapter. In particular, we have seen that the method of sieves is especially well suited to solve the problem of *excess simplicity* (as it allows for an unbounded infinite dimensional parameter space Θ). Furthermore, we have seen that the principle of indirect inference offers the possibility to deal with complex models and thus solves the problem of *excessive complexity*. Since we wish to solve both problems simultaneously, the task ahead of us consists (quite naturally) of combining the *method of sieves* with the principle of *indirect inference*. The resulting estimator shall be called a *semi-non parametric indirect inference* (SNPII) estimator.

Some properties of the SNPII estimator are worth noting immediately. First, in accordance with the method of sieves, the SNPII estimator will allow the parameter space to be unbounded and infinite dimensional. However, unlike the classical sieve estimator, criterion functions need not be analytically tractable. Second, in the spirit of indirect inference, the estimator will rely on the use of auxiliary statistics. However, unlike traditional indirect inference estimators, simulated data can be 'drawn' from a misspecified parametric model. In particular, in applications, data must only be simulated from the family of probability distributions indexed by the parameter $\boldsymbol{\theta}$ on the finite dimensional sieve Θ_T that approximates Θ.

Below, we provide an econometric motivation for the SNPII estimator. We use this example to relate the theory reviewed in this chapter with the estimation of models derived from economic theory.

Econometric Motivation

Consider a discrete-time version of the simple *Ramsey–Cass–Koopmans* model, the origins of which can be traced back to the seminal work of young mathematician

Frank Ramsey (1928). For illustrative purposes, our interpretation shall be limited in scope to the economics of a single farm. The structure of the problem is however shared by a large class of models whose influence is pervasive in various fields of economics ranging from Macroeconomics to Microeconomics and Empirical Finance.

Consider a corn producing farmer living in isolation. In any given year $t \in \mathbb{N}$, production of corn y_t, is a function of two variables. The stock of corn seeds k_t used for sowing, and a measure of the various (latent) exogenous productivity conditions z_t (e.g. meteorological events) affecting corn production. The relation between output and production factors is given by a *production function* $y_t = f(k_t, z_t)$. Every year the farmer must decide how much of the year's corn production y_t to consume, denoted c_t, and how much to reserve for next year's plantation k_{t+1}. The farmer's behavior is thus subject to a dynamic constraint of the form $k_{t+1} = f(k_t, z_t) - c_t$. Preferences over alternative streams of consumption $\{c_t\}_{t\in\mathbb{N}}$ are described by a utility function $U_t(c_1, c_2, ...) = \sum_{s=t}^{\infty} \beta^{s-t} u(c_s)$ where $u(c_t)$ denotes the *instantaneous utility* derived from consumption of c_t and β is a *time-preference parameter*. The description is complete by letting the exogenous *total factor productivity* (TFP) variable z_t exhibit dynamics described according to $z_t = g(z_{t-1}) + \epsilon_t$ where $\{\epsilon_t\}_{t\in\mathbb{N}}$ is an *iid* random sequence. Our agent's decisions are finally modeled as the solution to a nonlinear dynamic stochastic constrained optimization problem,

$$\max_{\{c_s\}_{s=t}^{\infty}} E_t \left[\sum_{s=t}^{\infty} \beta^{s-t} u(c_s) \right] \quad , \quad \text{s.t.} \quad k_{t+1} = f(k_t, z_t) - c_t \, , \quad z_t = g(z_{t-1}) + \epsilon_t. \tag{1.5}$$

Under sufficient smoothness assumptions the farmer's behavior can be described by a system of dynamic first-order conditions that includes both *(i)* a *consumption Euler equation* of the form, $u'(c_t) = \beta E[f'_k(k_{t+1}, z_{t+1}) u'(c_{t+1})]$, and *(ii)* the dynamic constraints postulated in (1.5). Depending on specific assumptions on how agents form expectations, the first-order conditions can be turned into a dynamic system of equations that is ultimately the focus of econometric analysis.

At this point, researchers will typically proceed by parameterizing the unknown functions u, f and g. Most importantly, the assumption will be made that the true *utility*, *production* and *TFP* functions can be represented as $u(c_t; \boldsymbol{\theta}_u)$, $f(k_t, z_t; \boldsymbol{\theta}_f)$ and $g(z_t; \boldsymbol{\theta}_g)$ for some vector of parameters $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_f, \boldsymbol{\theta}_g) \in \mathbb{R}^p$, $p \in \mathbb{N}$. Interested then lies in conducting inference on the *true* parameters and on the associated *true utility*, *production* and *TFP* functions.

The principle of indirect inference reviewed in Section 1.5 offers a way to proceed with the econometric analysis of such a model. Furthermore, the indirect inference estimator will not be disturbed by the presence of unobserved variables such as z_t and the potentially complicated dynamic structure that pose problems to the use of classical statistical techniques. When attempting to apply the principle of indirect inference researchers will however be faced with the important problem of *incorrect*

model specification.

Remark 1.6.1. *Typically, $(\boldsymbol{\theta}_u, \boldsymbol{\theta}_f, \boldsymbol{\theta}_g)$ is a vector of parameters in $\mathbb{R}^p$ for some small $p \in \mathbb{N}$. As a result, the functions u, f and g are assumed to belong to very restrictive classes of functions. This makes correct specification axioms very restrictive.*

Conducting statistical inference on parametric models under the influence of an axiom of correct specification is often hard to justify. Present economic theory can only provide us with a simple and stylized representation of what is potentially an immensely complex Data Generating Process (DGP). As mentioned in Sections 1.2 and 1.5, statistical inference in the absence of classical axioms of correct specification has long been available. Important econometric problems might however occur. First, pseudo-true parameters are not necessarily unique (see Chapter 5). Second, pseudo-true parameters are typically time-varying, and hence not structural. Finally, the deep economically meaningful interpretation of parameter estimates (in line with the underlying economic theory) is unavailable in the absence of an axiom of correct specification.

Remark 1.6.2. *Misspecification can cause pseudo-true parameters $\boldsymbol{\theta}_0^*$ to deviate considerably from $\boldsymbol{\theta}_0$ (White 1980b and Gourieroux et al. (1984)) thus being difficult to accept in the light of economic theory. This is especially worrying since the economic interpretation of parameter estimates is frequently used as an important indicator of the model's credibility.*

Quantities of interest such as estimates of the *output elasticity of capital*, *marginal rates of transformation* or the steady-state *Arrow-Pratt coefficient of relative risk aversion* (APC) are essentially void of any meaning without the influence of an axiom of correct specification. For example, $\widehat{APC} = -c_{ss}u''(c_{ss}, \hat{\boldsymbol{\theta}}_u)/u'(c_{ss}, \hat{\boldsymbol{\theta}}_u)$ is meaningless if $\hat{\boldsymbol{\theta}}_u$ does not correspond to the estimate of a *true* parameter.

The widely used *CRRA* utility function $u(c_t; \theta_u) = c_t^{1-\theta_u}/(1-\theta_u)$, for some positive scalar $\theta_u \neq 1$, might provide convenient estimates of this quantity since $\widehat{\text{APC}} = \hat{\theta}_u$. However, in the likely event of incorrect specification, $\hat{\theta}_u$ constitutes an estimate of a *pseudo-true* quantity that does not possess the intended economic interpretation.

Remark 1.6.3. *While algebraically tractable models might offer valuable analytical insight and elegant theoretical descriptions of economic activity, they have little credibility in a statistical context requiring the strict satisfaction of an axiom of correct specification.*

As argued in Section 1.2, allowing for a large infinite dimensional parameter space is a solution that attributes considerably more generality to the model at hand.

Indeed, interest in conducting statistical inference on unknown parameters lying on infinite dimensional spaces has gained popularity often as means of avoiding the restrictiveness of parametric models and the undesirable consequences of incorrect specification.

As mentioned before, the solution we shall study in the following chapters consists of combining the *principle of indirect inference* and the *method of sieves* in order to obtain sound statistical inference on complex nonlinear dynamic models featuring unobserved variables. This should allow us to obtain economically meaningful results that relate to the underlying economic theory.

Remark 1.6.4. *A simple example that illustrates the idea of SNPII estimation consists of parameterizing the utility function u according to,*[24]

$$u(c_t;\boldsymbol{\theta}_u) = \sum_{k=0}^{K_T} \theta_{u,k} c_t^k \quad , \quad \textit{with} \quad K_T \to \infty \quad \textit{as} \quad T \to \infty.$$

Under appropriate regularity conditions such a formulation would allow us to consistently estimate any continuous utility function (Debreu's theorem in Debreu (1959, p.56) shows that utility functions are continuous under very mild conditions on preferences).[25]

[24] The use of truncated power series as approximation devices in dynamic models is inappropriate for several reasons. We ignore this detail for now.

[25] Methods for imposing curvature restrictions such as monotonicity and concavity both locally and globally can be found e.g. in Diewert and Wales (1987) and Gallant and Golub (1984).

Chapter 2

Semi-Nonparametric Indirect Inference

This chapter introduces the Semi-Nonparametric Indirect Inference (SNPII) estimator in its most general form and provides a first account of its properties. In particular, we will discuss below the main results of consistency, convergence rate and asymptotic distribution of the SNPII estimator. While the consistency of the SNPII estimator is obtained as a special case of existing consistency theorems for sieve extremum estimators, the convergence rate and asymptotic distribution theorems are entirely new to sieve estimation. These two theorems generalize existing results that apply only to unrestricted extremum estimators (without sieves). Furthermore, they apply to a large class of smooth sieve estimators and thus constitute an addition to the general theory of sieve extremum estimation. The results in this chapter, in conjunction with those of Chapter 3, allow also for considerable more generality on both the nature of the estimator and the properties of the data generating process.

Unfortunately, a considerable number of technical details can easily cloud the ideas and arguments presented here. Indeed, the theory of sieve estimation seems quite prone to entangle itself on numerous inessential mathematical curiosities. This might lead one to believe that a certain proof is of a more complex nature than it really is. For this reason, this chapter will avoid dealing with such technicalities and take a more superficial and intuitive approach to the theory of SNPII estimation. We leave the technical precision and notational rigor to Chapter 3.

In essence, some notation will remain purposely simplistic and some arguments will appeal more to intuition than to a rigorous logical derivation. This will naturally come with some cost. It is possible that the reader might feel that a number of questions are left unanswered. I shall try to contain as much as possible such undesirable bi-products of the somewhat superficial treatment. In any case, Chapter 3 should dissipate any remaining doubts. There, the treatment of the SNPII

theory is more rigorous, the notation is precise and the arguments are laid down in their entirety.

In what follows, Section 2.1 introduces the SNPII estimator. Section 2.2 addresses its consistency. Section 2.3 presents a general theorem on the convergence rate of smooth extremum sieve estimators. Section 2.4 delivers a theorem for their asymptotic distribution. Finally, Section 2.5 discusses the assumptions used in the convergence and asymptotic distribution theorems.

2.1 Basic Formulation

The SNPII estimator $\hat{\boldsymbol{\theta}}_T$ belongs to the class of sieve extremum estimators introduced in Section 1.3. As such, it is generally described as the minimizer of a criterion function Q_T over a sequence of sieves $\{\Theta_T\}$, subsets of the parameter space Θ. For convenience let us restate the distinctive form of an *exact sieve extremum estimator*,

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\Theta_T} Q_T(\boldsymbol{\theta}). \tag{2.1}$$

A more general counterpart of this estimator, that contains (2.1) as a special case, is the *approximate sieve extremum estimator*,

$$Q_T(\hat{\boldsymbol{\theta}}_T) \;\leq\; \inf_{\boldsymbol{\theta}\in\Theta_T} Q_T(\boldsymbol{\theta}) + O_p(\eta_T) \tag{2.2}$$

for some $\eta_T \to 0$ as $T \to \infty$. Regardless of its formulation as in (2.1) or (2.2), a sieve estimator is called an SNPII estimator when its criterion function Q_T takes the special form of an indirect inference criterion function,

$$Q_T(\boldsymbol{\theta}) = \mu\Big(\hat{\boldsymbol{\beta}}_T \,,\; \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) \tag{2.3}$$

where, just as in Section 1.5, μ is some distance function (from now on called *criterion divergence*), $\hat{\boldsymbol{\beta}}_T$ is an auxiliary estimator obtained from observed data $\mathbf{x_T} := \mathbf{x}_1, ..., \mathbf{x}_T$, and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ denotes an average of S auxiliary estimators $1/S \sum_{s=1}^{S} \tilde{\boldsymbol{\beta}}_{T,s}(\boldsymbol{\theta})$ obtained from S streams of simulated data $\tilde{\mathbf{x}}^s_{\mathbf{T}}(\boldsymbol{\theta}) := \tilde{\mathbf{x}}^s_1(\boldsymbol{\theta}), ..., \tilde{\mathbf{x}}^s_T(\boldsymbol{\theta})$, $s = 1, ..., S$. The characterizing feature that makes the SNPII estimator distinct from other sieve extremum estimators is hence the form of its criterion function.

As pointed out in Section 1.5, any successful indirect inference estimator must rely on well-chosen 'informative' auxiliary estimators. In particular, in comparison to Θ, the auxiliary parameter space $\mathcal{B}$ should be sufficiently 'rich'. When Θ is a complex infinite-dimensional space, it becomes hard to ascertain what an 'appropriate' auxiliary parameter space $\mathcal{B}$ might be. A formal discussion of these problems is deferred to Chapter 3. Here we proceed intuitively.

Suppose that we make use of a single parametric auxiliary estimator, i.e. suppose that $\mathcal{B} \subseteq \mathbb{R}^q$, for some $q \in \mathbb{N}$. Then, one immediately suspects that $\mathcal{B}$ might not

be rich enough and consequently fail to be informative about a parameter $\boldsymbol{\theta}_0$ lying on an infinite dimensional space Θ. In particular, one suspects that the binding function $\boldsymbol{\beta}^*$ (see Section 1.5) will fail to be injective, yielding $\boldsymbol{\theta}_0$ unidentified.

There are several ways of avoiding this problem. One possible solution consists of choosing an auxiliary estimator that also takes values in an infinite dimensional space $\mathcal{B}$ that seems rich enough. In particular, one can make use of nonparametric auxiliary estimators as e.g. in Billio and Monfort (2003) and Nickl and Pötscher (2009).[1] As we shall see, another solution consists of using multiple (potentially infinitely many) parametric auxiliary estimators.

Remark 2.1.1. *An infinite number of parametric auxiliary estimators might turn out to provide a rich enough set of information for indirect inference to be successful. More generally, one might even choose to use infinitely many nonparametric or sieve auxiliary estimators, thus arriving at an auxiliary parameter space that is most certainly quite rich.*

Regardless of the specific choice that is made, this discussion should alert us for the need of a more general setting of SNPII estimation. In particular, a setting that allows for the use of infinitely many auxiliary statistics. Such a setting is embodied in the following definition of the criterion function,

$$Q_T(\boldsymbol{\theta}) = \mu_T\Big(\hat{\boldsymbol{\beta}}_T\ ,\ \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) \tag{2.4}$$

where μ_T is some distance function (that now depends on T), $\hat{\boldsymbol{\beta}}_T$ is an infinite vector of auxiliary estimators, $(\hat{\boldsymbol{\beta}}_T^1,\ \hat{\boldsymbol{\beta}}_T^2,\ ...)$ obtained from observed data $\mathbf{x_T}$, and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ is an infinite vector $(\tilde{\boldsymbol{\beta}}_{T,S}^1(\boldsymbol{\theta}),\ \tilde{\boldsymbol{\beta}}_{T,S}^2(\boldsymbol{\theta}),\ ...)$ of averages of auxiliary estimators $\tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta})$ obtained from S streams of simulated data $\tilde{\mathbf{x}}_{\mathbf{T}}^s(\boldsymbol{\theta})$, $s = 1, ..., S$ (recall Section 1.5).

Remark 2.1.2. *Following Gourieroux et al. (1993), auxiliary estimators might consist e.g. of extremum estimators on spaces $\mathcal{B}_i$ with criterion functions $\mathcal{Q}_T^i$,*

$$\hat{\boldsymbol{\beta}}_T^i = \arg\min_{\boldsymbol{\beta}_i \in \mathcal{B}_i} \mathcal{Q}_T^i(\mathbf{x_T}, \boldsymbol{\beta}_i) \quad \textit{and} \quad \tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\beta}_i \in \mathcal{B}_i} \mathcal{Q}_T^i(\tilde{\mathbf{x}}_{\mathbf{T}}^s(\boldsymbol{\theta}), \boldsymbol{\beta}_i) \quad \textit{for every } i \in \mathbb{N}.$$

A simpler formulation consists e.g. of auxiliary estimators obtained as,

$$\hat{\boldsymbol{\beta}}_T^i = \frac{1}{T}\sum_{t=1}^{T} \mathcal{Q}_T^i(\mathbf{x}_t) \quad \textit{and} \quad \tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{Q}_T^i(\tilde{\mathbf{x}}_t^s(\boldsymbol{\theta})) \quad \textit{for every} \quad i \in \mathbb{N}.$$

In (2.4), the need for a *criterion divergence* μ_T that depends on T is justified by very practical reasons. In particular, note that (in applications) it is simply impossible for us to make use of infinitely many auxiliary estimators. Hence, we need a *criterion divergence* μ_T that 'gives positive weight' only to a finite subset of

[1] We could even choose to use other sieve estimators as auxiliary statistics.

the infinite vectors $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$. When T is finite, this allows us to effectively 'neglect' all the auxiliary statistics that are 'given zero weight' by μ_T. At the same time, we can construct the sequence $\{\mu_T\}$ in such a way as to give positive weight to all statistics, asymptotically. More details about the construction of such a sequence are given in Chapter 3.

What is important to retain at this point is that the simpler criterion function in (2.3) is just a special case of the more general one in (2.4). Hence, apart restrictions on the generality of Θ, all results derived for the more general case apply also to the simpler one. From now on, we thus let the *exact SNPII estimator* be defined as,

$$\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\Theta_T} \mu_T\Big(\hat{\boldsymbol{\beta}}_T \,,\, \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big),$$

and the *approximate SNPII estimator* be defined as,

$$\mu_T\Big(\hat{\boldsymbol{\beta}}_T \,,\, \tilde{\boldsymbol{\beta}}_{T,S}(\hat{\boldsymbol{\theta}}_T)\Big) \;\leq\; \inf_{\boldsymbol{\theta}\in\Theta_T} \mu_T\Big(\hat{\boldsymbol{\beta}}_T \,,\, \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) + O_p(\eta_T), \tag{2.5}$$

for some $\eta_T \to 0$. In Chapter 3, a certain formulation of this estimator is shown to be measurable, consistent, $\sqrt{T}$-convergent and asymptotically normal.

2.2 Consistency Structure

Let $\boldsymbol{\theta}_0$ be the unique minimizer of Q_∞. Then, $\boldsymbol{\theta}_0$ is sometimes called the *true parameter* and consistency of $\hat{\boldsymbol{\theta}}_T$ is defined as,[2]

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \xrightarrow{p} 0 \quad \text{as} \quad T \to \infty,$$

where $\|\cdot\|$ is a norm on Θ. Since the SNPII estimator is just a special case of the general sieve extremum estimator introduced in Section 1.3, we can make use of the known conditions for consistency in the hope of obtaining $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \xrightarrow{p} 0$. Recall from Lemma 1.3.1 that the conditions for consistency are the following.

List 1. (Consistency Assumptions)

1. *Sieves Θ_T are compact and satisfy $\Theta_T \subseteq \Theta_{T+1}$ for every T.*
2. *The sequence of sieves $\Theta_1, \Theta_2, \ldots$ is dense in Θ;*
3. *$\sup_{\boldsymbol{\theta}\in\Theta_T} |Q_T(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| = o_p(1)$ as $T \to \infty$;*
4. *The limit criterion function Q_∞ is continuous on Θ;*
5. *The parameter $\boldsymbol{\theta}_0$ is the identifiably unique minimizer of Q_∞.*

[2]The 'true' parameter is always w.r.t. an underlying distribution. Again, $\xrightarrow{p}$ denotes convergence in probability and $\|\cdot\|$ a norm on any given vector space.

Conditions 1 and 2 are regularity conditions that are trivially satisfied by an appropriate definition of the sieves Θ_T and the parameter space Θ. In this respect the SNPII estimator is no different from all the other sieve estimators. Appropriate results can thus be found throughout the literature on sieve estimators. We are thus left we the task of showing that conditions 3, 4 and 5 hold for the special case of a criterion function Q_T given by (2.3). These conditions are discussed in Chapter 3.

2.3 Convergence Rate

Unfortunately, the theory of convergence rates for sieve extremum estimators is much less developed than that of consistency. In particular, there are still no available theorems that apply to the entire class of sieve extremum estimators.[3] Hence, in this section, we cannot simply refer to any existing set of conditions. We will have to devise them by ourselves. Luckily, much is already done. As we shall see, our task below will consist fundamentally of adapting van der Vaart's theorem discussed in Section 1.2 (Lemma 1.2.1) into something that is useful for sieve extremum estimators.

Recall that van der Vaart's theorem derived the convergence rate and asymptotic distribution of sufficiently smooth extremum estimators on infinite dimensional spaces. Now, observing Lemma 1.2.1, it might seem that it applies also to sieve estimators. After all, the parameter space is allowed to be infinite dimensional and it is possible that the introduction of sieves does not have any influence on the result. We are thus led to ask the questions: *is van der Vaart's theorem applicable in this case? Does the theorem require any adaptation to deal with sieves?* The answer is quite simply, *no* and *yes*, respectively. The argument goes as follows.

Let Θ be an infinite dimensional vector space. Recall from Section 1.2 that the system of partial derivatives of Q_T is denoted ∇Q_T and seen as an infinite system of estimating equations. From the continuous invertibility of $\nabla^2 Q_\infty$ at $\boldsymbol{\theta}_0$ and appropriate smoothness conditions, van der Vaart concludes essentially that, for some $c > 0$,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \;\leq\; c \cdot \|\nabla Q_T(\boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0)\| \tag{2.6}$$

and hence that $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| = O_p(r_T^{-1})$, for some real sequence $r_T \to \infty$, is implied by having,

$$\|\nabla Q_T(\boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0)\| = O_p(r_T^{-1}) \quad \text{as} \quad T \to \infty. \tag{2.7}$$

In the context of unrestricted estimation, the argument that $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| = O_p(r_T^{-1})$ follows from (2.7) is perfectly logical in its conclusion. In the presence of a sieve

[3] As pointed out in Section 1.3, some important results do exist however for sieve M-estimators.

estimator however, the convergence rate of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|$ cannot be deduced solely from the behavior of $\|\nabla Q_T(\boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0)\|$ in (2.6).

Remark 2.3.1. *In the presence of sieve restrictions, the behavior of the criterion function Q_T is no longer the sole determinant of the behavior of $\hat{\boldsymbol{\theta}}_T$ (which is also restricted by the sieves). This leads us to conclude that the theorem in van der Vaart (1995) needs some adaptation if we wish to apply it to sieve estimators.*

The adaptation (called for in the remark above) is described next. In sieve estimation, the sieves are precisely designed to restrict the behvaior of $\hat{\boldsymbol{\theta}}_T$. In essence, the following condition must be added to (2.6),

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \geq \|\pi_T(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\| \quad \forall\ T \in \mathbb{N},$$

where $\pi_T(\boldsymbol{\theta}_0)$ is the projection of $\boldsymbol{\theta}_0$ on Θ_T. This should solve our problem. As it turns out however, making explicit use of this condition in combination with (2.6) is more complicated than it seems at first and leads very easily to dead ends. Hence, instead of imposing this condition explicitly, we shall deal with it implicitly. The idea is to 'split' the convergence problem into two parts using the following inequality,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \ \leq\ \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| \ +\ \|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|, \tag{2.8}$$

where $\boldsymbol{\theta}_T^0 := \arg\min_{\boldsymbol{\theta}\in\Theta_T} Q_\infty(\boldsymbol{\theta})$ denotes the minimizer of Q_∞ over the sieve Θ_T. Then, the desired result follows naturally by showing that,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| = O_p(r_T^{-1}) \quad \text{and} \quad \|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = O(r_T^{-1}) \quad \text{as} \quad T \to \infty.$$

The first part, concerned with showing that $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| = O_p(r_T^{-1})$, takes place in a probabilistic setting of unrestricted convergence 'within' each sieve. This condition can be addressed in a setting similar to that of van der Vaart's, since within the sieves Θ_T, convergence of $\hat{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}_T^0$ is determined essentially by the appropriate convergence of ∇Q_T to ∇Q_∞. The second part, is concerned with showing $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = O(r_T^{-1})$ and takes place in a deterministic setting 'outside' the sieves. This condition can be obtained by imposing appropriate smoothness conditions on ∇Q_∞ and on the rate of expansion of the sieves.

Theorem 2.3.1 below obtains the desired r_T convergence rate of $\hat{\boldsymbol{\theta}}_T$. For simplicity, we assume that $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = O(r_T^{-1})$ holds true. This condition is quite easy to obtain and hence we defer its discussion to a later section. In this way we can focus first on the more complicated part which consists of providing sufficient conditions for $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| = O_p(r_T^{-1})$. For the sake of clarity, the theorem makes use of very general 'high-order' assumptions on the sieves and criterion functions. In Chapter 3 we show how to derive these assumptions from more primitive conditions that are

suitable for our SNPII estimator. Finally, note that from now on, we shall make use of a more 'refined' notation concerning derivatives in infinite dimensional spaces.

Let $\mathbb{S}_\Theta$ denote the system of basis vectors of Θ.[4] Until now, given a function f on Θ, we have implicitly assumed that $\nabla f(\boldsymbol{\theta})$ denotes the vector of derivatives of f at $\boldsymbol{\theta} \in \Theta$ in the direction of the vectors of $\mathbb{S}_\Theta$. Hence, a vector of *partial derivatives* at $\boldsymbol{\theta}$. This notation worked well until now because we were always interested in derivatives in the directions of $\mathbb{S}_\Theta$. With the introduction of sieves however, we will often be interested in analyzing derivatives in other directions.

Remark 2.3.2. *From now on, $\nabla f(\boldsymbol{\theta}', \boldsymbol{\theta})$ denotes the derivative of a map f at $\boldsymbol{\theta}'$ in the direction of $\boldsymbol{\theta}$. Sometimes we also write $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}')$ or $f^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}')$ instead of $\nabla f(\boldsymbol{\theta}', \boldsymbol{\theta})$. Accordingly, when $\mathbb{S}_\Theta$ is a system of basis vectors, the vector of partial derivatives at $\boldsymbol{\theta}$, previously denoted $\nabla f(\boldsymbol{\theta})$, is now denoted $\nabla f(\boldsymbol{\theta}, \mathbb{S}_\Theta)$, sometimes $\nabla_{\mathbb{S}_\Theta} f$ or $f^{\nabla_{\mathbb{S}_\Theta}}$.*

In accordance with the remark above, the 'system of estimating equations' is now denoted either $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta)$ or $\nabla_{\mathbb{S}_\Theta} Q_\infty$, and sometimes, $Q_\infty^{\nabla_{\mathbb{S}_\Theta}}$.[5] In finite samples, the sieve estimator makes use of a system $\nabla Q_T(\cdot, \mathbb{S}_{\Theta_T})$ with a finite number of 'estimating equations', i.e. the system of derivatives in the direction of the basis vectors $\mathbb{S}_{\Theta_T}$ of the finite-dimensional sieve Θ_T.

The interested reader is advised to read Section C.1 in Appendix C for more information on the notational convention applied henceforth concerning derivatives of operators on infinite dimensional spaces.

Theorem 2.3.1. (Convergence Rate of Sieve Extremum Estimator) *Let $\hat{\boldsymbol{\theta}}_T$ be a sieve extremum estimator as defined in (2.1) satisfying $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$. Let Q_T and Q_∞ be differentiable on Θ. Furthermore, let $\nabla_{\mathbb{S}_\Theta} Q_\infty$ be continuously differentiable in a neighborhood of $\boldsymbol{\theta}_0$. Suppose that $\hat{\boldsymbol{\theta}}_T$ satisfies an 'approximate' Z-estimator formulation $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$ and that sieve expansion rates ensure $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o(r_T^{-1})$. In addition suppose that the following smoothness condition holds true,*

$$r_T \left\| (\nabla_{\mathbb{S}_{\Theta_T}} Q_T - \nabla_{\mathbb{S}_\Theta} Q_\infty)(\hat{\boldsymbol{\theta}}_T) - (\nabla_{\mathbb{S}_{\Theta_T}} Q_T - \nabla_{\mathbb{S}_\Theta} Q_\infty)(\boldsymbol{\theta}_T^0) \right\| = o_p(1 + r_T \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|). \tag{2.9}$$

Finally, let the derivative of $Q_\infty^{\nabla_{\mathbb{S}_\Theta}}$, denoted $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}$, satisfy a continuous invertibility condition ensuring that for large enough T, there exists some $c > 0$ such that,

$$\left\| \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T, \boldsymbol{\theta}') \right\| \geq c \cdot \|\boldsymbol{\theta}'\| \text{ for every } \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0 \text{ and } \boldsymbol{\theta}' \in \operatorname{lin}(\Theta). \tag{2.10}$$

[4] As we shall see in Chapter 3 an appropriate basis for Θ is the *Schauder* basis.

[5] The use of three different ways of denoting the same element might seem unnecessarily complicated. As we shall see however, notations of the type $f^{\nabla_{\boldsymbol{\theta}}}$ are very useful in shortening otherwise very long expressions, especially when the direction plays an uninteresting role. On the other hand, notations of the type $\nabla f(\cdot, \boldsymbol{\theta})$ are very convenient when interest lies in derivations focusing on the directions of derivatives.

Then, if $\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_T^0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\| = O_p(r_T^{-1})$ *it follows that the smooth sieve extremum estimator* $\hat{\boldsymbol{\theta}}_T$ *satisfies* $r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| = O_p(1)$ *as* $T \to \infty$.

Proof. Notice first that $\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0) = 0$.[6] Now, by the continuous invertibility assumption in (2.10) and the assumption that $\boldsymbol{\theta}_T^0 \to \boldsymbol{\theta}_0$, we have that for large enough T, there exists a $c > 0$ such that,

$$\left\|\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0, \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0)\right\| \;\geq\; c \cdot \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|. \tag{2.11}$$

By noting that the continuous differentiability of $Q_\infty^{\nabla_{\mathbb{S}_\Theta}}$ on a neighborhood of $\boldsymbol{\theta}_0$ implies,

$$\left\|Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\hat{\boldsymbol{\theta}}_T) - Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0) - \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0, \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0)\right\| = o(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|)$$

as $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| \xrightarrow{p} 0$,[7] it follows immediately from (2.11) that,

$$\left\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\right\| \;\geq\; c \cdot \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| + o(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|)$$

Rearranging the inequality, multiplying both sides by r_T, and norm sub-additivity imply,

$$\begin{aligned} r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|(c + o(1)) \;&\leq\; r_T\left\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\right\| \\ &\leq\; r_T\left\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T)\right\| + r_T\left\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\right\| \end{aligned} \tag{2.12}$$

The differentiability of $\nabla_{\mathbb{S}_\Theta} Q_\infty$ on a neighborhood of $\boldsymbol{\theta}_0$ together with $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o(r_T^{-1})$ implies that $\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\| = o(r_T^{-1})$. Furthermore, since $\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)\| = o_p(r_T^{-1})$ holds by assumption, we can conclude that the right-hand-side of (2.12) satisfies,

$$\begin{aligned} r_T\left\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T)\| + r_T\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\right\| \;&\leq\; r_T\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)\| \\ &\qquad + r_T\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)\| + o(1) \\ &\leq\; r_T\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)\| + o_p(1) \end{aligned}$$

As a result, (2.12) can be re-written as,

$$r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|(c + o(1)) \;\leq\; r_T\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)\| + o_p(1) \tag{2.13}$$

Now, making use of the smoothness condition in (2.9), it follows from (2.13) that,

$$r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|(c + o(1)) \;\leq\; r_T\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_T^0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\| + o_p(1 + r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|).$$

[6] Clearly, this condition is not, in any way, affected by the introduction of sieves Recall also that it is unnecessarily restrictive to impose that $\boldsymbol{\theta}_0$ be the unique element of Θ satisfying $\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0) = 0$.

[7] $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| \xrightarrow{p} 0$ is implied by $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \xrightarrow{p} 0$ and $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| \xrightarrow{p} 0$.

Finally, since $\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_T^0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\| = O_p(r_T^{-1})$, it follows naturally that,

$$r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|\Big(c + o(1) - o_p(1)\Big) \leq O_p(1) \;\; \Leftrightarrow \;\; r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| \leq \frac{O_p(1)}{c + o_p(1)} = O_p(1).$$

The desired result is thus obtained, since as $T \to \infty$,

$$r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \;\leq\; r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| \;+\; r_T\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| \;=\; O_p(1) + O(1) = O_p(1).$$

□

Several conditions imposed in Theorem 2.3.1 are quite abstract and thus require further explanation. In Section 2.5 we discuss these conditions in more detail. For the sake of simplicity and universality of application, the discussion is still kept at a fairly high level of generality. Primitive conditions are introduced in Chapter 3.

Finally, before moving on to the next section, it is important to clarify an issue concerning the convergence rate of what is commonly called the '*score*' in parametric ML estimation. In applications, it is typically the case that $r_T\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\|$ is known to be $O_p(1)$. Just an in van der Vaart (1995) it is essentially this boundedness in probability that determines the convergence rates of extremum estimators; see Lemma 1.2.1. However, in Theorem 2.3.1 we have made use of the alternative somewhat ambiguous assumption that,

$$\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_T^0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\| = O_p(r_T^{-1}). \tag{2.14}$$

Luckily, Theorem 2.4.1 below introduces a smoothness condition under which (2.14) boils down precisely to the more familiar condition on the convergence rate of the 'score', i.e. on the rate at which $\|\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\|$ vanishes as $T \to \infty$.

2.4 Asymptotic Distribution

Let us now turn to the asymptotic distribution of $r_T\Big(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\Big)$. Again, it should be noted that the following theorem actually holds for the general class of sieve extremum estimators. As before, we make use of high-level assumptions which help in keeping the main argument quite simple and clear.

Theorem 2.4.1. (Asymptotic Distribution of Sieve Extremum Estimator) *Let the conditions of Theorem 2.3.1 hold. Assume further that,*

$$r_T\Big\|(\nabla_{\mathbb{S}_{\Theta_T}} Q_T - \nabla_{\mathbb{S}_\Theta} Q_\infty)(\boldsymbol{\theta}_T^0) - (\nabla_{\mathbb{S}_{\Theta_T}} Q_T - \nabla_{\mathbb{S}_\Theta} Q_\infty)(\boldsymbol{\theta}_0)\Big\| = o_p(1 + r_T\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \tag{2.15}$$

Then, if $r_T\Big(\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\Big) \xrightarrow{d} \mathbb{G}$ *for some stochastic process* $\mathbb{G}$*, it follows that,*

$$r_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} -\mathrm{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\big(\boldsymbol{\theta}_0, \cdot\big)\Big)(\mathbb{G}) \quad as \quad T \to \infty.^{8}$$

Proof. Following essentially the same argument as in Theorem 2.3.1, the smoothness condition (2.9) and the fact that $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$ and $\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0) = o(r_T^{-1})$ implies,

$$\begin{aligned} r_T\Big[\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\Big] &= r_T \nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) + o(1) \\ &= r_T\Big[\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)\Big] + o_p(1) \\ &= -r_T\Big[\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_T^0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\Big] \\ &\quad + o_p(1 + r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|) \end{aligned} \tag{2.16}$$

Now making use of the novel smoothness condition in (2.15) we obtain from (2.16),

$$\begin{aligned} r_T\Big[\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\Big] = &-r_T\Big[\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\Big] \\ &+ o_p(1 + r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|) + o(1 + r_T\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|). \end{aligned}$$

Finally, since $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o_p(r_T^{-1})$ and $r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\| = O_p(1)$ (derived in Theorem 2.3.1), it follows immediately that,

$$r_T\Big[\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\Big] = -r_T\Big[\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\Big] + o_p(1). \tag{2.17}$$

Now, the differentiability of $\nabla_{\mathbb{S}_\Theta} Q_\infty$ at $\boldsymbol{\theta}_0$ implies that,

$$\Big\|\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0) - \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)\Big\| = o(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|)$$

holds as $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|$ vanishes.[9] Furthermore, since $\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0) = 0$ and since $\nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0) = o(r_T^{-1})$ follows from differentiability of $\nabla_{\mathbb{S}_\Theta} Q_\infty$ and $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o(r_T^{-1})$, we obtain,

$$\begin{aligned} r_T \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) &= r_T\Big[\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\Big] + o(r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|) \\ &= r_T\Big[\nabla_{\mathbb{S}_\Theta} Q_\infty(\hat{\boldsymbol{\theta}}_T) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_T^0)\Big] + o_p(1) + o_p(r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|) \end{aligned} \tag{2.18}$$

This implies together with (2.17) and $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| = O_p(r_T^{-1})$ that,

$$\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\big(\boldsymbol{\theta}_0, r_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)\big) = -r_T\Big[\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\Big] + o_p(1)$$

[8] Recall again that $\mathrm{inv}(f)$ denotes the inverse of an operator f. Also, $\xrightarrow{d}$ denotes convergence in distribution. The *stochastic process* $\mathbb{G}$ denotes a random element on an infinite dimensional space.

[9] $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \xrightarrow{p} 0$ follows from the assumed consistency of $\hat{\boldsymbol{\theta}}_T$.

and equivalently, by the continuous invertibility of $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}$ at $\boldsymbol{\theta}_0$,

$$r_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = -\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\big(\boldsymbol{\theta}_0, \cdot\big)\Big)\Big(r_T\big[\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\big]\Big) + o_p(1).$$

The desired result follows from $r_T\Big(\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta}_0) - \nabla_{\mathbb{S}_\Theta} Q_\infty(\boldsymbol{\theta}_0)\Big) \xrightarrow{d} \mathbb{G}$ and an application of the continuous mapping theorem,

$$r_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} -\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\big(\boldsymbol{\theta}_0, \cdot\big)\Big)(\mathbb{G}) \quad \text{as} \quad T \to \infty.$$

□

In the context of indirect inference estimation, the asymptotic distribution $\mathbb{G}$ of the 'score' in Theorem 2.4.1 above is obtained from the asymptotic distribution of the auxiliary statistics. This is a common feature of indirect inference estimators; see e.g. Gourieroux et al. (1993) and Gourieroux and Monfort (1996). There is however one novelty introduced here that complicates this inferential strategy.

When SNPII estimation is performed with an infinite vector of auxiliary statistics, then the asymptotic distribution $\mathbb{G}$ above must be derived from the asymptotic distribution of infinitely many auxiliary estimators. In practice, this is impossible. Fortunately, Chapter 3 reveals that an approximation can be devised. The idea will be to make use only of the asymptotic distribution of the finite number of auxiliary statistics used in estimation (those given positive weight by the criterion divergence μ_T). Loosely speaking, this will amount to derive the asymptotic distribution of a classical indirect inference estimator (as in Gourieroux et al. (1993)), yet regarding it only as an approximation to the actual asymptotic distribution of the SNPII estimator derived above.

In essence, by requiring the correct specification axiom to hold only asymptotically, the SNPII estimator allows us to claim important generality.[10] However, for some SNPII estimators (like the one considered in Chapter 3), this comes at a price. For those SNPII estimators that make use of infinitely many auxiliary statistics, the asymptotic distribution derived in Theorem 2.4.1 above is not analytically tractable. Instead, only an approximation is available. As we shall see, the quality of this approximation depends on the complexity of the parameter space, the choice of sieves, and the choice (and number) of auxiliary estimators. Most importantly, Chapter 3 provides the theoretical foundations that justify this approximation.

[10]Strictly speaking, it is not correct to talk about correct specification 'holding only asymptotically' as the parameter space is fixed. Intuitively however, this idea turns out to be quite useful.

2.5 Intermediate Conditions

This section takes some first steps towards the verification of the assumptions imposed in the theorems above. Some conditions introduced in those theorems were designed essentially to make the proofs simple. However, due to their abstract nature, it is hard to know whether they hold in practice. Our objective here is to 'decompose' the more abstract assumptions into a number of simpler sufficient conditions that we shall call *intermediate conditions*. By doing this, we will effectively pave the way for the more detailed discussion of sufficient *primitive conditions* in Chapter 3. List 2 below presents the *high-level* more abstract assumptions we shall discuss in this section.

List 2. (High-Level Assumptions)

1. *Z-estimator formulation* $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$,
2. *Convergence rate of the constrained minimizer* $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o_p(r_T^{-1})$,
3. *Smoothness of the criterion process,*
   $r_T\|(\nabla_{\mathbb{S}_{\Theta_T}} Q_T - \nabla_{\mathbb{S}_\Theta} Q_\infty)(\hat{\boldsymbol{\theta}}_T) - (\nabla_{\mathbb{S}_{\Theta_T}} Q_T - \nabla_{\mathbb{S}_\Theta} Q_\infty)(\boldsymbol{\theta}_T^0)\| = o_p(1 + r_T\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|)$,

The intermediate conditions that we shall use to obtain these high-level assumptions above are the following.

List 3. (Sufficient Intermediate Conditions)

1. *Approximation error in (2.2) satisfies* $\eta_T = o(r_T^{-1})$,
2. *Expansion rate of sieves satisfies* $\sup_{\boldsymbol{\theta}\in\Theta} \|\pi_T(\boldsymbol{\theta}) - \boldsymbol{\theta}\| = o_p(r_T^{-1})$,
3. *For every* $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$, *there exists some* $\bar{c} > 0$ *and large enough* $T^* \in \mathbb{N}$ *such that,*
   $\left|\nabla Q_\infty^{\nabla_\theta}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}'')\right| \geq \bar{c} \cdot \|\boldsymbol{\theta}''\|$ *holds for all* $T > T^*$ *and* $(\boldsymbol{\theta}, \boldsymbol{\theta}'') \in \text{lin}\Theta \times \text{lin}\Theta$,[11]
4. *For every* $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$, *there exists some* $\bar{c} > 0$ *and large enough* $T^* \in \mathbb{N}$ *such that,*
   $\left|\nabla Q_T^{\nabla_\theta}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}'')\right| \geq \bar{c} \cdot \|\boldsymbol{\theta}''\|$ *holds a.s. for all* $T > T^*$ *and* $(\boldsymbol{\theta}, \boldsymbol{\theta}'') \in \text{lin}\Theta \times \text{lin}\Theta$,
5. $\left|Q_\infty(\boldsymbol{\theta}_T + \boldsymbol{\theta}'_T) - Q_\infty(\boldsymbol{\theta}_T) - \nabla Q_\infty(\boldsymbol{\theta}_T, \boldsymbol{\theta}'_T)\right| = o(\|\boldsymbol{\theta}'_T\|)$
   holds as $\|\boldsymbol{\theta}'_T\| \to 0$ *for every* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$,
6. $\left|Q_T(\boldsymbol{\theta}_T + \boldsymbol{\theta}'_T) - Q_T(\boldsymbol{\theta}_T) - \nabla Q_T(\boldsymbol{\theta}_T, \boldsymbol{\theta}'_T)\right| = o(\|\boldsymbol{\theta}'_T\|)$
   holds a.s. as $\|\boldsymbol{\theta}'_T\| \to 0$ *for every* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$,[12]

[11] lin(Θ) or linΘ denote the linear span of the vector space Θ.

[12] Here a.s. stands for *almost surely*.

7. $\left|Q_T^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T+\boldsymbol{\theta}''_T)-Q_T^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T)-\nabla Q_T^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T,\boldsymbol{\theta}''_T)\right|=o(\|\boldsymbol{\theta}''_T\|)$
holds a.s. as $\|\boldsymbol{\theta}''_T\|\to 0$ *for every* $\boldsymbol{\theta}'_T\to\boldsymbol{\theta}_0$ *and every* $\boldsymbol{\theta}_T\to\boldsymbol{\theta}\in\text{lin}\Theta$,

8. $\left|Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T+\boldsymbol{\theta}''_T)-Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T)-\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T,\boldsymbol{\theta}''_T)\right|=o(\|\boldsymbol{\theta}''_T\|)$
holds as $\|\boldsymbol{\theta}''_T\|\to 0$ *for every* $\boldsymbol{\theta}'_T\to\boldsymbol{\theta}_0$ *and every* $\boldsymbol{\theta}_T\to\boldsymbol{\theta}\in\text{lin}\Theta$.

As we shall see in Chapter 3, condition 1 in List 3 above is obtained essentially by defining appropriately the approximate extremum estimator in (2.2). Condition 2 is obtained, under appropriate regularity conditions, by appealing to results stemming from the Approximation Theory literature.

Conditions 3 and 4 are essentially continuous invertibility conditions. In particular, Chapter 3 reveals that Conditions 3 and 4 follow, under appropriate regularity and compact convergence conditions, from the continuous invertibility of $\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0,\cdot)$ for every direction $\boldsymbol{\theta}\in\text{lin}\Theta$. Appendix B provides the main results.

Conditions 5, 6, 7 and 8 are essentially smoothness conditions related to Gateaux, Hadamard and Fréchet differentiability concepts. Unfortunately however, none of these can be directly applied. In this thesis, we are thus forced to introduce novel (albeit related) smoothness concepts that are better suited for the task at hand. In practice, we need to 'extend' the traditional concepts of differentiability on infinite dimensional spaces to hold over sequences of functions and uniformly over differentiability points and directions.

The interested reader is encouraged to take a look at Sections C.2 and C.3, in Appendix C, which introduce *(i)* the notion of *uniform differentiability of the first, second and third kinds*, *(ii)* the concept of *Hadamard sequence* of operators, and finally *(iii)* the important smoothness concept of (*uniform*) *Hadamard equidifferentiability* (*of the first, second and third kinds*) of a sequence of operators. Chapter 3 will analyze these concepts in considerable more detail. Appendix C provides a characterization of classes of functions satisfying these novel smoothness concepts. Let us now analyze the conditions of List 2, one by one.

Z-Estimator Formulation of $\hat{\boldsymbol{\theta}}_T$

The Z-estimator formulation of $\hat{\boldsymbol{\theta}}_T$ (condition 1 of List 2), i.e. the formulation of $\hat{\boldsymbol{\theta}}_T$ as an estimator satisfying $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T)=o_p(r_T^{-1})$, was simply assumed to hold in Theorem 2.3.1. However, since $\hat{\boldsymbol{\theta}}_T$ is a constrained minimizer of Q_T, there is no *a priori* reason to suppose that this condition holds in every conceivable setting. Intermediate conditions must thus be devised to ensure that such a characterization of the sieve extremum estimator $\hat{\boldsymbol{\theta}}_T$ holds true.

The following argument is substantially simplified by further 'partitioning' the distance $\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_0\|$ into 'smaller components'. Let $\boldsymbol{\theta}^*_T$ denote a minimizer of Q_T over

the sieve Θ_T and $\boldsymbol{\theta}_T^{**}$ denote a minimizer of Q_T over the entire Θ,

$$\boldsymbol{\theta}_T^* \in \arg\min_{\boldsymbol{\theta}\in\Theta_T} Q_T(\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\theta}_T^{**} \in \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\boldsymbol{\theta}).$$

Then, for every $T \in \mathbb{N}$, the following simple inequality holds true,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\| \;\leq\; \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| \;+\; \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \;+\; \|\boldsymbol{\theta}_T^{**} - \boldsymbol{\theta}_T^0\| \;+\; \|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|. \tag{2.19}$$

As we shall see, this condition can be obtained with more or less effort, depending on the exact definition of the estimator $\hat{\boldsymbol{\theta}}_T$ and the nature of the sieves. For example, if $\hat{\boldsymbol{\theta}}_T$ is an *exact extremum sieve estimator* as defined in (2.1), then $\hat{\boldsymbol{\theta}}_T = \boldsymbol{\theta}_T^*$ holds by construction. This is not the case however if $\hat{\boldsymbol{\theta}}_T$ is an *approximate extremum sieve estimator* as given by (2.2). Furthermore, if the sieves Θ_T are *purely dimensional w.r.t.* Q_T (Definition A.82 and Remarks A.83 and A.84), then it follows immediately that $\hat{\boldsymbol{\theta}}_T = \boldsymbol{\theta}_T^* = \boldsymbol{\theta}_T^{**}$ and hence $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = 0$ for every T. In this case, the so-called *Z-estimator formulation of* $\hat{\boldsymbol{\theta}}_T$ is readily available. If one of these conditions fail however, then $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$ can still be obtained under additional smoothness and invertibility conditions in List 3. Below, we analyze some important sufficient conditions. More details can be found in Chapter 3

Suppose that the smoothness condition 7 in List 3 holds,

$$\left| Q_T^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T + \boldsymbol{\theta}''_T) - Q_T^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T) - \nabla Q_T^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}''_T) \right| = o(\|\boldsymbol{\theta}''_T\|) \tag{2.20}$$

a.s. as $\|\boldsymbol{\theta}''_T\| \to 0$ for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$ and every $\boldsymbol{\theta}_T \to \boldsymbol{\theta} \in \text{lin}\Theta$. Then, (2.20) holds trivially for fixed $\boldsymbol{\theta}_T = \boldsymbol{\theta}$ for every $\boldsymbol{\theta} \in \mathbb{S}_\Theta$. As we shall see in Chapter 3 under appropriate topological conditions (involving the use of Tychonoff's topology on appropriate sets), it then follows that,

$$\left\| Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}'_T + \boldsymbol{\theta}''_T) - Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}'_T) - \nabla Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}''_T) \right\| = o(\|\boldsymbol{\theta}''_T\|) \tag{2.21}$$

holds also a.s. as $\|\boldsymbol{\theta}''_T\| \to 0$ for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$. As a result, using the fact that $\nabla_{\mathbb{S}_{\Theta_T}} Q(\boldsymbol{\theta}_T^{**}) = 0$, in Chapter 3 we conclude that, under appropriate topological regularity conditions, $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$ follows essentially from having $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^{**}\| = o_p(r_T^{-1})$.

Now, as pointed out in the discussion preceding Theorem 2.3.1, we must be careful in analyzing the convergence rate of quantities of the type $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^{**}\|$ since the convergence of $\hat{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}_T^{**}$ depends on both the behavior of the criterion function Q_T and the sieves Θ_T. Fortunately, by using the inequality,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^{**}\| \;\leq\; \|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| \;+\; \|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \tag{2.22}$$

we can once again 'split' the argument into the convergence rate of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\|$ (which occurs 'within sieves' and depends essentially on Q_T), and the convergence rate of

$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|$ (which depends essentially on the behavior of the sieves). Indeed, the desired characterization of $\hat{\boldsymbol{\theta}}_T$ as an approximate Z-estimator, i.e. as an estimator satisfying $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$, can thus be obtained by showing that $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| = o_p(r_T^{-1})$ and $\|\boldsymbol{\theta}_T^{**} - \boldsymbol{\theta}_T^*\| = o_p(r_T^{-1})$. We now turn to this task.

Clearly, the convergence rate of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\|$ is fundamentally related to the term η_T in (2.2) which turns $\hat{\boldsymbol{\theta}}_T$ into an approximate minimizer of Q_T on the sieve Θ_T. In particular, we expect that, under appropriate conditions, having $\eta_T = o(r_T^{-1})$ (condition 1 in List 3) should yield the desired result $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| = o_p(r_T^{-1})$.

Let the condition 4 in List 3 hold. In particular, suppose that for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$, there exists some $\bar{c} > 0$ and large enough $T^* \in \mathbb{N}$ such that,

$$\left|\nabla Q_T^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}'')\right| \;\geq\; \bar{c} \cdot \|\boldsymbol{\theta}''\| \tag{2.23}$$

holds a.s. for all $T > T^*$ and $(\boldsymbol{\theta}, \boldsymbol{\theta}'') \in \text{lin}\Theta \times \text{lin}\Theta$.

Then, since $\boldsymbol{\theta}_T^* \xrightarrow{p} \boldsymbol{\theta}_0$, there exists some $\bar{c} > 0$ and large enough T^* such that,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| \;\leq\; \bar{c}\left|\nabla Q_T^{\nabla_{\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*}}(\boldsymbol{\theta}_T^*, \hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*)\right|$$

holds for all $T > T^*$. Together with the smoothness condition (2.20) (condition 7 in List 3) this implies that,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| \;\leq\; \bar{c}\left|Q_T^{\nabla_{\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*}}(\hat{\boldsymbol{\theta}}_T) - Q_T^{\nabla_{\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*}}(\boldsymbol{\theta}_T^*)\right| + o(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\|)$$

as $T \to \infty$. Now, making use of condition 6 in List 3 and noting that both $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_T^* \xrightarrow{p} \boldsymbol{\theta}_0$, we obtain,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| \;\leq\; \bar{c}\left|Q_T(\hat{\boldsymbol{\theta}}_T) - Q_T(\boldsymbol{\theta}_T^*)\right| + \bar{c}\left|Q_T(\hat{\boldsymbol{\theta}}_T) - Q_T(\boldsymbol{\theta}_T^*)\right| + o(\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\|).$$

as $T \to \infty$ with probability tending to one. Finally, $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\| = o_p(r_T^{-1})$ follows by an appropriate convergence of η_T in (2.2) (condition 1 in List 3) and by noting that $\left|Q_T(\hat{\boldsymbol{\theta}}_T) - Q_T(\boldsymbol{\theta}_T^*)\right| \leq O_p(\eta_T)$. In particular, as $T \to \infty$,

$$\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^*\|(1 + o(1)) \;\leq\; 2\bar{c}O_p(\eta_T) = o_p(r_T^{-1}).$$

A similar result can be obtained for $\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|$. Notice first that the convergence rate of $\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|$ is restricted by the inequality,

$$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \;\geq\; \|\pi_T(\boldsymbol{\theta}_T^{**}) - \boldsymbol{\theta}_T^{**}\| \tag{2.24}$$

imposed by the sieves, where $\pi_T(\boldsymbol{\theta}_T^{**})$ denotes the projection of $\boldsymbol{\theta}_T^{**}$ on Θ_T. However, since $\sup_{\boldsymbol{\theta}\in\Theta} \|\pi_T(\boldsymbol{\theta}) - \boldsymbol{\theta}\| \geq \|\pi_T(\boldsymbol{\theta}_T^{**}) - \boldsymbol{\theta}_T^{**}\|$, condition 2 of List 3, $\sup_{\boldsymbol{\theta}\in\Theta} \|\pi_T(\boldsymbol{\theta}) - \boldsymbol{\theta}\| = o_p(r_T^{-1})$ ensures that $\|\pi_T(\boldsymbol{\theta}_T^{**}) - \boldsymbol{\theta}_T^{**}\| = o_p(r_T^{-1})$. As such, obtaining $\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| = o_p(r_T^{-1})$ is not an impossibility.

Now, since $\boldsymbol{\theta}_T^{**} \to \boldsymbol{\theta}_0$, we can make once again use of the invertibility condition (2.23) (condition 4 in List 3) to conclude that $\exists\, \bar{c} > 0$ and $T^* \in \mathbb{N}$ such that,

$$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \;\leq\; \bar{c}\Big|\nabla Q_T^{\nabla\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}}(\boldsymbol{\theta}_T^{**}, \boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**})\Big|$$

holds a.s. for $T > T^*$. By the smoothness condition in (2.20) (condition 7 in List 3) it then follows that,

$$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \;\leq\; \bar{c}\Big|Q_T^{\nabla\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}}(\boldsymbol{\theta}_T^*) - Q_T^{\nabla\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}}(\boldsymbol{\theta}_T^{**})\Big| + o(\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|)$$

as $T \to \infty$. Now, using the fact that $Q_T^{\nabla\boldsymbol{\theta}}(\boldsymbol{\theta}_T^{**}) = 0$ holds by construction for every $\boldsymbol{\theta} \in \text{lin}\Theta$, and condition 6 in List 3, it holds as before that,

$$\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \;\leq\; \bar{c}\Big|Q_T(\boldsymbol{\theta}_T^*) - Q_T(\boldsymbol{\theta}_T^{**})\Big| + o(\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|)$$

Finally, since it holds by definition that $\Big|Q_T(\boldsymbol{\theta}_T^*) - Q_T(\boldsymbol{\theta}_T^{**})\Big| \;\leq\; \bar{c}\Big|Q_T(\pi_T(\boldsymbol{\theta}_T^{**})) - Q_T(\boldsymbol{\theta}_T^{**})\Big|$, it is shown in Chapter 3 that under appropriate compact convergence and topological regularity conditions,

$$\begin{aligned}\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\| \;&\leq\; \bar{c}\Big|Q_T(\pi_T(\boldsymbol{\theta}_T^{**})) - Q_T(\boldsymbol{\theta}_T^{**})\Big| + o(\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|)\\ &=\; O(\|\pi_T(\boldsymbol{\theta}_T^{**}) - \boldsymbol{\theta}_T^{**}\|) + o(\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|)\end{aligned}$$

as $T \to \infty$. Together with (2.24), this implies that $\|\boldsymbol{\theta}_T^* - \boldsymbol{\theta}_T^{**}\|$ has the same asymptotic convergence rate as $\|\pi_T(\boldsymbol{\theta}_T^{**}) - \boldsymbol{\theta}_T^{**}\|$ which is $o_p(r_T^{-1})$. In light of (2.22) and the preceding argument it thus follows that $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\hat{\boldsymbol{\theta}}_T) = o_p(r_T^{-1})$.

Convergence Rate of $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|$

Recall that the simple inequality in (2.8) showed that an upper bound on the convergence rate of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|$ can be derived by analyzing separately the convergence rates of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_T^0\|$ and $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|$. Recall also that in Theorem 2.3.1 we have simply imposed that $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o(r_T^{-1})$ as an assumption (condition 2 of List 2). However, in applications this rate of convergence is likely to be unknown from the outset. What is typically known by the researcher is the convergence rate of $\|\pi_T(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|$ not that of $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|$.

Indeed, the convergence rate of $\|\pi_T(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|$ can be exogenously determined by the researcher by appealing to an appropriate design of the sieves and results available in the Approximation Theory literature.[13] Let us now turn to the task of providing sufficient conditions for $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| = o(r_T^{-1})$ to hold.

[13] In practice, stronger results of the convergence rate of $\sup_{\boldsymbol{\theta}\in\Theta} \|\pi_k(\boldsymbol{\theta}) - \boldsymbol{\theta}\|$ as a function of a variable k are typically available.

Just as in the previous section, we make use of an invertibility condition that offers an upper bound to $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|$. Here we turn to condition 3 in List 3. As a result, for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$, there exists some $\bar{c} > 0$ and large enough $T^* \in \mathbb{N}$ such that,[14]

$$\left|\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}'')\right| \;\geq\; \bar{c} \cdot \|\boldsymbol{\theta}''\| \tag{2.25}$$

holds for all $T > T^*$ and $(\boldsymbol{\theta}, \boldsymbol{\theta}'') \in \text{lin}\Theta \times \text{lin}\Theta$,

We will also make use of two smoothness conditions of List 3 that fall over Q_∞ and its directional derivatives. First, condition 5 ensures that,

$$\left|Q_\infty(\boldsymbol{\theta}_T + \boldsymbol{\theta}'_T) - Q_\infty(\boldsymbol{\theta}_T) - \nabla Q_\infty(\boldsymbol{\theta}_T, \boldsymbol{\theta}'_T)\right| = o(\|\boldsymbol{\theta}'_T\|) \tag{2.26}$$

holds as $\|\boldsymbol{\theta}'_T\| \to 0$ for every $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. Second, condition 8 ensures that,

$$\left|Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T + \boldsymbol{\theta}''_T) - Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T) - \nabla Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}(\boldsymbol{\theta}'_T, \boldsymbol{\theta}''_T)\right| = o(\|\boldsymbol{\theta}''_T\|) \tag{2.27}$$

holds as $\|\boldsymbol{\theta}''_T\| \to 0$ for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}_0$ and every $\boldsymbol{\theta}_T \to \boldsymbol{\theta} \in \text{lin}\Theta$.

Now, since $\boldsymbol{\theta}_T^0 \to \boldsymbol{\theta}_0$ (implied by the denseness of the sieves on Θ), there exists some $\bar{c} > 0$ and large enough $T^* \in \mathbb{N}$ such that,

$$\begin{aligned}
\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| &\leq \bar{c}\left|\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\right| \\
&\leq \bar{c}\left|Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_T^0) - Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_0)\right| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \\
&= \bar{c}\left|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\right| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \\
&= \bar{c}\left|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\right| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|)
\end{aligned} \tag{2.28}$$

where the first inequality follows from the invertibility condition (2.25), the second inequality from the smoothness assumption on the directional derivatives (2.27), the first equality holds by definition, and the second equality by noting that $\nabla Q_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0$ for every $\boldsymbol{\theta}$. As a result, we have that,

$$\begin{aligned}
\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\| &\leq \bar{c}\left|Q_\infty(\boldsymbol{\theta}_T^0) - Q_\infty(\boldsymbol{\theta}_0)\right| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \\
&\leq \bar{c}\left|Q_\infty(\pi_T(\boldsymbol{\theta}_0)) - Q_\infty(\boldsymbol{\theta}_0)\right| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \\
&= O(\|\pi_T(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|) + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|)
\end{aligned}$$

as $T \to \infty$, where the first inequality follows from (2.28) and the smoothness condition (2.26), the second inequality is implied by the definition of $\boldsymbol{\theta}_T^0$ and the last equality follows again by the smoothness of Q_∞. Finally, from $\|\pi_T(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\| = o(r_T^{-1})$

[14] Under an appropriate product topology structure, Chapter 3 shows that this invertibility condition implies also condition (2.10) used in Theorem 2.3.1.

we obtain,

$$\begin{aligned}&\|\boldsymbol{\theta}_T^0-\boldsymbol{\theta}_0\|(1+o(1))=O(\|\pi_T(\boldsymbol{\theta}_0)-\boldsymbol{\theta}_0\|)\\&\Leftrightarrow\|\boldsymbol{\theta}_T^0-\boldsymbol{\theta}_0\|=\frac{1}{(1+o(1))}O(o(r_T^{-1}))=o(r_T^{-1}).\end{aligned}$$

Smoothness of $\nabla_{\mathbb{S}_{\Theta_T}}Q_T-\nabla_{\mathbb{S}_\Theta}Q_\infty$

Finally, we turn to the last item (condition 3) in List 2 of *high-level* assumptions.

$$r_T\|(\nabla_{\mathbb{S}_{\Theta_T}}Q_T-\nabla Q_\infty)(\hat{\boldsymbol{\theta}}_T)-(\nabla_{\mathbb{S}_{\Theta_T}}Q_T-\nabla Q_\infty)(\boldsymbol{\theta}_T^0)\|=o_p(1+r_T\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0\|).\quad(2.29)$$

This smoothness condition, used in Theorem 2.3.1, is very similar to that in van der Vaart (1995) and captures well the main smoothness requirement used in Theorem 2.3.1 to obtain the appropriate convergence rate of the extremum sieve estimator $\hat{\boldsymbol{\theta}}_T$. Below, we verify that this assumption can be obtained by imposing more intelligible intermediate conditions. Fortunately, the intermediate conditions we search for are suggested by the use of a simple inequality,

$$\begin{aligned}&r_T\Big\|(\nabla_{\mathbb{S}_{\Theta_T}}Q_T-\nabla_{\mathbb{S}_\Theta}Q_\infty)(\hat{\boldsymbol{\theta}}_T)-(\nabla_{\mathbb{S}_{\Theta_T}}Q_T-\nabla_{\mathbb{S}_\Theta}Q_\infty)(\boldsymbol{\theta}_T^0)\Big\|\\&\quad\le\ r_T\Big\|\nabla_{\mathbb{S}_\Theta}Q_\infty(\hat{\boldsymbol{\theta}}_T)-\nabla_{\mathbb{S}_\Theta}Q_\infty(\boldsymbol{\theta}_T^0)-\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0)\Big\|\\&\qquad+r_T\Big\|\nabla_{\mathbb{S}_{\Theta_T}}Q_T(\hat{\boldsymbol{\theta}}_T)-\nabla_{\mathbb{S}_{\Theta_T}}Q_T(\boldsymbol{\theta}_T^0)-\nabla Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0)\Big\|\\&\qquad+\Big\|\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,r_T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0))-\nabla Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,r_T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0))\Big\|.\end{aligned}\quad(2.30)$$

The immediate conclusion to be drawn from (2.30) is that (2.29) is implied by having,

$$r_T\Big\|\nabla_{\mathbb{S}_\Theta}Q_\infty(\hat{\boldsymbol{\theta}}_T)-\nabla_{\mathbb{S}_\Theta}Q_\infty(\boldsymbol{\theta}_T^0)-\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0)\Big\|=o_p(1+r_T\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0\|),\quad(2.31)$$

$$r_T\Big\|\nabla_{\mathbb{S}_{\Theta_T}}Q_T(\hat{\boldsymbol{\theta}}_T)-\nabla_{\mathbb{S}_{\Theta_T}}Q_T(\boldsymbol{\theta}_T^0)-\nabla Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0)\Big\|=o_p(1+r_T\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0\|),\quad(2.32)$$

and

$$\Big\|\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,r_T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0))-\nabla Q_T^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,r_T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0))\Big\|=o_p(1+r_T\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0\|).\quad(2.33)$$

Since $\boldsymbol{\theta}_T^0\to\boldsymbol{\theta}_0$ holds by the denseness of the sieves $\{\Theta_T\}$ on Θ and $\|\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0\|\xrightarrow{p}0$ holds by consistency of $\hat{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}_0$, the condition (2.31) is implied by condition 7 of List 3. Also, by the same argument, (2.32) is implied by condition 5 of List 3. As previously mentioned, the conditions of List 5 are presented and characterized in Appendix C.

Finally, (2.33) has been left out of List 3, as it does not require any new special differentiability theory or invertibility condition. In fact, in Chapter 3 this condition is obtained as a bi-product of a *delta method* for Hadamard operators, by appealing to the tightness of $r_T(\hat{\boldsymbol{\theta}}_T-\boldsymbol{\theta}_T^0)$ on a separable space.

2.6 Final Remarks

This chapter introduced the SNPII estimator in its most general form and presented the main results of consistency, convergence rate and asymptotic distribution.

While the SNPII framework provided a useful starting point, the novel theorems of convergence and asymptotic distribution presented in this chapter are applicable to the general class of sieve extremum estimators.

As mentioned in the introduction, these theorems seem to fill an important gap in the literature. The usefulness of these results is well appreciated in conjunction with those of Appendices B and C which provide an understanding of the nature of the intermediate conditions in List 3.

The high-level conditions used in this chapter were useful in retaining generality while keeping the proofs simple. Furthermore, it allowed us to separate the *general theory*, which applies to smooth sieve extremum estimators, from the *special theory* in Chapter 3 that applies only to the case of SNPII estimators. This separation is lost (or becomes at least difficult to recognize) in the following chapter. Chapter 3 will now address more carefully the main ideas reviewed here, and provide a more rigorous treatment of the theory covered in this chapter.

Chapter 3

A $\sqrt{T}$-Consistent and Asymptotically Gaussian SNPII Estimator

As argued in Chapter 1, econometric analysis of probability models featuring latent or unobserved variables and complex nonlinear dynamics is often challenging as such features might invalidate the use of classical estimators. This is true in particular for some well known M and Z estimators. For example, under such conditions, likelihood functions are likely to be intractable, yielding Maximum Likelihood (ML) techniques inappropriate. Likewise, moment conditions might be hard to derive analytically, restricting the availability of Method of Moments (MM) estimators. To avoid such problems, simulation-based counterparts of these estimators are often considered. In this chapter we establish the properties of a sieve estimator that relies on the unifying principle of *indirect inference* introduced in Gourieroux et al. (1993) and Smith (1993).[1]

The present chapter proposes a sieve extremum estimator for semi-nonparametric models that relies on an infinite vector of parametric auxiliary statistics through the principle of indirect inference. The estimator is shown to be $\sqrt{T}$-consistent and asymptotically Gaussian under general regularity conditions. The data is allowed to exhibit heterogeneous and dependent behavior. Furthermore, in the tradition of indirect inference, these results apply to a large class of complex dynamic models with unobserved variables, including those yielding an estimator with no closed form algebraic representation or featuring a criterion function which is intractable or infeasible, even on appropriately chosen compact finite-dimensional sieves.

In what follows, Section 3.1 introduces some preliminary assumptions on the DGP and parameter spaces. At the cost of some repetition, Section 3.2 defines

[1]Most simulation estimators are a special case of the *indirect inference* estimator. See Gourieroux and Monfort (1996) for details.

again the SNPII estimator. This time however, in considerable more detail. Sections 3.3 and 3.4 establish measurability and consistency, respectively. Section 3.5 derives the $\sqrt{T}$-convergence rate and asymptotic Gaussianity. Section 3.6 discusses the possibility of conducting inference with an approximation of the asymptotic distribution. Heterogeneity and dependence issues are briefly addressed in Section 3.7. Section 3.8 gives some remarks on the uniform convergence of auxiliary estimators. Section 3.9 discusses the surprising $\sqrt{T}$-convergence result (in norm) in the context of known optimal convergence theorems for estimators on infinite dimensional spaces. In particular, it explains the apparent contradiction between the convergence rate obtained here and some theorems in the literature of optimal convergence rates. Section 3.10 concludes. Finally, Section 3.11 contains the proofs of the main theorems and propositions.

Before moving on, a word on notation. Throughout, we let $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of natural, integer and real numbers respectively. Given a set $\mathbb{A}$, we let $\mathcal{T}_{\mathbb{A}}$ denote a topology on $\mathbb{A}$. Given a topological space $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$, we let $\mathfrak{B}(\mathbb{A})$ denote the Borel σ-algebra generated by $\mathcal{T}_{\mathbb{A}}$, and denote the closure of $\mathbb{A}$ by $\text{cl}(\mathbb{A})$. A metric on $\mathbb{A}$ is denoted $\delta_{\mathbb{A}}$. When $\mathbb{A}$ is a vector space, then $\|\cdot\|_{\mathbb{A}}$ denotes a norm on $\mathbb{A}$. If $\mathbb{A}$ is a subset of a vector space, then $\text{lin}(\mathbb{A})$ denotes the linear span of $\mathbb{A}$. Given a metric space $(\mathbb{A}, \delta_{\mathbb{A}})$ we let $S(a_0, \epsilon)$ denote an open ball of radius $\epsilon > 0$ centered at $a_0 \in \mathbb{A}$. Occasionally, we might also adopt the notation $S_{a_0}(\epsilon)$, i.e. $S_{a_0}(\epsilon) := \{a \in \mathbb{A} : \delta_{\mathbb{A}}(a_0, a) < \epsilon\}$ and let $S^c_{a_0}(\epsilon)$ be its complement in $\mathbb{A}$. For any index set $\mathbb{I}$ and a collection of sets $\mathbb{A}_i$, $i \in \mathbb{I}$, we let $\times_{i\in\mathbb{I}}\mathbb{A}_i$ denote the Cartesian product of the sets $\mathbb{A}_i$. Projections operators are denoted $\pi_i : \mathbb{A} \to \mathbb{A}_i$. Given two topological spaces $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ we let $\mathbb{C}(\mathbb{A}, \mathbb{B})$ denote the space of continuous functions mapping from $\mathbb{A}$ into $\mathbb{B}$. If $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ are topological vector spaces then $\mathbb{L}(\mathbb{A}, \mathbb{B})$ denotes the space of bounded linear operators from $\mathbb{A}$ into $\mathbb{B}$. Finally, given suitably defined random variables, $\xrightarrow{d}$, $\xrightarrow{p}$ and $\xrightarrow{a.s.}$ denote convergence in distribution, probability and almost surely, respectively.

3.1 Data Generating Process and Parameter Spaces

Observed data consists of a T-sequence $\mathbf{x_T}(\omega) := \{\mathbf{x}_t(\omega)\}_{t=1}^T$ of points in $\mathbb{R}^{n_\mathbf{x}}$, $(T, n_\mathbf{x}) \in \mathbb{N} \times \mathbb{N}$, a subset of the realized path of an $n_\mathbf{x}$-variate stochastic sequence $\mathbf{x}(\omega) = \{\mathbf{x}_t(\omega), t \in \mathbb{Z}\}$ for some element ω of the event space Ω of a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\mathcal{F}$ denotes a σ-algebra defined on Ω and $\mathbb{P}$ a probability measure on $\mathcal{F}$.[2] The random sequence $\mathbf{x}$ is thus an $\mathcal{F}/\mathfrak{B}(\mathbb{R}^{n_\mathbf{x}}_\infty)$-measurable mapping $\mathbf{x} : \Omega \to \mathbb{R}^{n_\mathbf{x}}_\infty$ taking values in the Cartesian product of infinite copies of $\mathbb{R}^{n_\mathbf{x}}$, denoted

[2]Given a measurable space $(\mathcal{A}, \mathfrak{B}(\mathcal{A}))$, a measurable map $f : \Omega \to \mathcal{A}$ and some $A \in \mathfrak{B}(\mathcal{A})$, we shall often write $\mathbb{P}(A)$ instead of $\mathbb{P}(\{\omega \in \Omega : a(\omega) \in A\})$ when there is no risk of ambiguity.

$\mathbb{R}^{n_{\mathbf{x}}}_{\infty} := \times_{t=-\infty}^{t=\infty}\mathbb{R}^{n_{\mathbf{x}}}$. The stochastic sequence $\mathbf{x}(\omega)$ lives on the space $(\mathbb{R}^{n_{\mathbf{x}}}_{\infty}, \mathfrak{B}^{n_{\mathbf{x}}}_{\infty}, \mathrm{D}_0)$ where the induced probability measure (p.m.) D_0 is naturally defined over the elements of the Borel σ-algebra $\mathfrak{B}^{n_{\mathbf{x}}}_{\infty} := \mathfrak{B}(\mathbb{R}^{n_{\mathbf{x}}}_{\infty})$ generated by the finite dimensional product cylinders of $\mathbb{R}^{n_{\mathbf{x}}}_{\infty}$. In general, $\mathbf{x}$ is allowed to exhibit various common forms of time-dependence and heterogeneity. This has been noted in Section 1.5 when the *indirect inference* estimator was introduced and it is further discussed in Section 3.7. The precise level of generality depends on the choice of auxiliary estimators. Indeed, all that matters is that the data confers auxiliary estimators with adequate properties.

In general, the model of interest consists of a family $\mathcal{D}_\Theta$ of p.m.s $\mathrm{D}(\boldsymbol{\theta})$ defined on $\mathfrak{B}^{n_{\mathbf{x}}}_{\infty}$. We let the elements of this family be indexed by a possibly infinite-dimensional parameter $\boldsymbol{\theta} \in \Theta$, so that $\mathcal{D}_\Theta = \{\mathrm{D}(\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$.[3] The metric space (Θ, δ_Θ), called *parameter space*, is assumed to possess some general properties of interest. These properties are laid down in the following assumption and are satisfied by various function spaces, e.g. the space of continuous functions C, Lebesgue spaces L^p for $1 \leq p < \infty$, Hölder spaces $\mathcal{H}^p$, Sobolev spaces $W^{k,p}$ for $p < \infty$, and others. Non-separable spaces such as the spaces of bounded functions L^∞ or bounded sequences ℓ^∞ are however excluded.

Assumption 3.1.1. *The parameter space (Θ, δ_Θ) is a complete, separable, measurable metric space with Borel σ-algebra $\mathfrak{B}(\Theta)$ generated by the topology $\mathcal{T}_\Theta$ induced by the metric δ_Θ on Θ.*

Subsets of Θ, called *sieves*, will be indexed by $T \in \mathbb{N}$ and denoted $\Theta_T \subseteq \Theta\ \forall T \in \mathbb{N}$. The sieves are typically designed to possess desirable features (e.g. compactness) that are especially convenient for working with extremum estimators (recall Section 1.1). A mild form of correct specification is also assumed. Namely, that $\exists\, \boldsymbol{\theta}_0 \in \Theta$ such that $\mathrm{D}(\boldsymbol{\theta}_0) = \mathrm{D}_0$. However, we allow for the possibility that $\boldsymbol{\theta}_0 \notin \Theta_T\ \forall T \in \mathbb{N}$, requiring only that the sequence of sieves be increasing and dense on Θ. This is a distinct characteristic of the method of sieves.

Assumption 3.1.2. *The sieves $\{\Theta_T\}_{T\in\mathbb{N}}$ are non-empty compact subsets of Θ satisfying $\Theta_T \subseteq \Theta_{T+1} \subseteq \Theta\ \forall T \in \mathbb{N}$ and $\mathrm{cl}\big(\bigcup_{T\in\mathbb{N}}\Theta_T\big) \supseteq \Theta$. Furthermore, $\exists\, \boldsymbol{\theta}_0 \in \Theta$: $\mathrm{D}(\boldsymbol{\theta}_0) = \mathrm{D}_0$, i.e. $\mathrm{D}_0 \in \mathcal{D}_\Theta$.*

Indirect inference on $\boldsymbol{\theta}_0$ is to be conducted under the assumption that it is possible to "draw" from the distribution $\mathrm{D}(\boldsymbol{\theta})$ for every $\boldsymbol{\theta}$ lying on well chosen (possibly finite dimensional) subsets $\Theta_T \subseteq \Theta$, $\forall T \in \mathbb{N}$. In other words, it must be possible to obtain T-period subsets $\tilde{\mathbf{x}}_{\mathbf{T}}(\boldsymbol{\theta}, \omega) := \{\tilde{\mathbf{x}}_t(\boldsymbol{\theta}, \omega)\}_{t=1}^{T}$ of the realized path

[3] Let $\mathcal{D}$ denote the set of all probability measures on $\mathfrak{B}^{n_{\mathbf{x}}}_{\infty}$, then, by definition, the subset $\mathcal{D}_\Theta \subseteq \mathcal{D}$ is the image of Θ under $\mathrm{D} : \Theta \to \mathcal{D}$ with $\mathrm{D}(\boldsymbol{\theta}) : \mathfrak{B}^{n_{\mathbf{x}}}_{\infty} \to [0, 1]$ for every $\boldsymbol{\theta} \in \Theta$.

of the stochastic sequence $\tilde{\mathbf{x}}(\boldsymbol{\theta}) : \Omega \to \mathbb{R}^{n_{\mathbf{x}}}_{\infty}$, living in $(\mathbb{R}^{n_{\mathbf{x}}}_{\infty}, \mathfrak{B}^{n_{\mathbf{x}}}_{\infty}, \mathrm{D}(\boldsymbol{\theta}))$, for every $\boldsymbol{\theta} \in \Theta_T$, $\forall T \in \mathbb{N}$. This seems hardly restrictive in practice, see e.g. Gourieroux and Monfort (1996). Note in particular that we do not require the ability to "draw" from D_0 since it is possible that $\mathrm{D}_0 \notin \mathcal{D}_{\Theta_T}\ \forall T \in \mathbb{N}$, where $\mathcal{D}_{\Theta_T} := \{\mathrm{D}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta_T\}$, even though $\mathrm{D}_0 \in \mathcal{D}_{\Theta}$. In this sense, statistical inference is conducted using a sequence of possibly misspecified models $\mathcal{D}_{\Theta_T} \subseteq \mathcal{D}\ \forall T \in \mathbb{N}$. When considering several draws of T-period sequences $\tilde{\mathbf{x}}_{\mathbf{T}}(\boldsymbol{\theta})$ from $\mathrm{D}(\boldsymbol{\theta})$, these shall be indexed by $s \in \{1, ..., S\}$, $S \in \mathbb{N}$ and denoted $\tilde{\mathbf{x}}^s_{\mathbf{T}}(\boldsymbol{\theta})$.

Finally, we define also a topological vector space $(\mathcal{B}, \mathcal{T}_{\mathcal{B}})$ called the *auxiliary parameter space. Indirect inference* on elements of Θ shall be conducted "through" inference on elements of $\mathcal{B}$. The auxiliary space $\mathcal{B}$ is obtained as the Cartesian product of a collection of *auxiliary factor spaces* $\mathcal{B}_i$, $i \in \mathbb{N}$, where $\mathbb{N}$ is a countable index set. We require that an appropriate topology be defined on the product space $\mathcal{B} = \times_{i\in\mathbb{N}}\mathcal{B}_i$, namely the Tychonoff's product topology. This ensures continuity of the projection maps $\pi_i : \mathcal{B} \to \mathcal{B}_i\ \forall i \in \mathbb{N}$ (Lemma A.15), a property which is crucial for the theory of *semi-nonparametric indirect inference*.[4] We shall often require $\mathcal{B}$ to be equipped with a metric $\delta_{\mathcal{B}}$. As such $(\mathcal{B}, \mathcal{T}_{\mathcal{B}})$ is assumed to be metrizable, and hence also Hausdorff (Lemma A.9).[5] Clearly, it is imposed from the outset that the metric $\delta_{\mathcal{B}}$ on $\mathcal{B}$ be a product metric inducing the product topology $\mathcal{T}_{\mathcal{B}}$ on $\mathcal{B}$. Examples of product metrics inducing the desired topology on countable product spaces are,

$$\delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}') = \sum_{i=1}^{\infty} \frac{1}{2^i} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}'_i\|_{\mathcal{B}_i}}{1 + \|\boldsymbol{\beta}_i - \boldsymbol{\beta}'_i\|_{\mathcal{B}_i}} \quad \text{and} \quad \delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}') = \sup_{i\in\mathbb{N}} \frac{1}{i} \frac{\|\boldsymbol{\beta}_i - \boldsymbol{\beta}'_i\|_{\mathcal{B}_i}}{1 + \|\boldsymbol{\beta}_i - \boldsymbol{\beta}'_i\|_{\mathcal{B}_i}}, \quad (3.1)$$

for every $(\boldsymbol{\beta}, \boldsymbol{\beta}') \in \mathcal{B} \times \mathcal{B}$ (Lemma A.22) where $\|\cdot\|_{\mathcal{B}_i}$ are norms on the factor vector spaces $\mathcal{B}_i$. Finally, note that measurability statements involving $(\mathcal{B}, \mathcal{T}_{\mathcal{B}})$ are made w.r.t. the Borel σ-algebra $\mathfrak{B}(\mathcal{B})$ generated by $\mathcal{T}_{\mathcal{B}}$.

Assumption 3.1.3. *The auxiliary parameter space $(\mathcal{B}, \delta_{\mathcal{B}})$ is a measurable metric space, a countable Cartesian product $\mathcal{B} := \times_{i\in\mathbb{N}}\mathcal{B}_i$ of subsets of complete separable normed topological vector spaces $(\mathcal{B}_i, \mathcal{T}_{\mathcal{B}_i})$ equipped with norms $\|\cdot\|_{\mathcal{B}_i}$ for every $i \in \mathbb{N}$. The product space $(\mathcal{B}, \delta_{\mathcal{B}})$ is equipped with a metric $\delta_{\mathcal{B}} : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ inducing Tychonoff's topology $\mathcal{T}_{\mathcal{B}}$ on $\mathcal{B}$ and a Borel σ-algebra $\mathfrak{B}(\mathcal{B})$ generated by $\mathcal{T}_{\mathcal{B}}$.*[6]

[4] An immediate consequence is that convergence of a sequence $\{b_T\}_{T\in\mathbb{N}}$ on $\mathcal{B}$ implies (and is implied by) the convergence of the projection sequences $\{\pi_i(b_T)\}_{T\in\mathbb{N}}$ on $\mathcal{B}_i\ \forall i \in \mathbb{N}$ (Corollary A.16). This in turn implies that continuity of operators f mapping from any topological space $\mathbb{A}$ into $\mathcal{B}$ holds if and only if $\pi_i \circ f : \mathbb{A} \to \mathcal{B}_i$ is continuous $\forall i \in \mathbb{N}$ (Lemma A.18). Moreover, compactness of subsets of $\mathcal{B}^* = \times_{i\in\mathbb{N}}\mathcal{B}^*_i \subseteq \mathcal{B}$ follows from compactness $\forall \mathcal{B}^*_i$ (Lemma A.19).

[5] This can be obtained by having $\mathcal{B}_i$ be regular and second countable for every $i \in \mathbb{N}$ (Lemmas A.7, A.20 and A.21).

[6] We do not require the metric on the linear space to be a norm. The latter requires homogeneity

Note that Assumption 3.1.3 implies that $\mathcal{B}$ is separable (Lemma A.21) and second countable (Lemma A.25). By Lemma A.24 this implies that $\mathfrak{B}(\mathcal{B}) = \otimes_{i\in\mathbb{N}}\mathfrak{B}(\mathcal{B}_i)$ where $\otimes_{i\in\mathbb{N}}\mathfrak{B}(\mathcal{B}_i)$ denotes the product σ-algebra. These algebras are thus used interchangeably. Most importantly, the projection mappings $\pi_i : \mathcal{B} \to B_i$ are $\mathfrak{B}(\mathcal{B})/\mathfrak{B}(\mathcal{B}_i)$-measurable $\forall i \in \mathbb{N}$ (Corollary A.17).[7] In Section 3.5 below, the factor spaces $\mathcal{B}_i$ are assumed to satisfy $\mathcal{B}_i \subseteq \mathbb{R}^n$, $n \in \mathbb{N}$ for every $i \in \mathbb{N}$, so that $\sqrt{T}$-convergence and asymptotic normality results are specific to SNPII estimators making use of infinitely many parametric auxiliary estimators.

3.2 The SNPII Estimator

Before attempting to analyze the properties of the SNPII estimator in detail, let us first define it more rigorously. Recall from Chapter 2 that the SNPII estimator is a map $\hat{\boldsymbol{\theta}}_{T,S} : \Omega \to \Theta_T$ satisfying, for fixed $S \in \mathbb{N}$,[8]

$$\hat{\boldsymbol{\theta}}_{T,S} \in \arg\min_{\boldsymbol{\theta}\in\Theta_T} Q_{T,S}(\boldsymbol{\theta}) \quad \text{a.s. } \forall T \in \mathbb{N}, \tag{3.2}$$

where $Q_{T,S} : \Omega \times \Theta \to \mathbb{R}$ is called the *criterion function* and $\Theta_T \subseteq \Theta$ as imposed by Assumption 3.1.2. In what follows, conditions shall be imposed on the criterion functions $Q_{T,S} : \Theta \times \Omega \to \mathbb{R}$ and the sieves Θ_T so as to guarantee that the arg min set exists and that $Q_{T,S}$ converges in some appropriate sense to a limit deterministic criterion function $Q_\infty : \Theta \to \mathbb{R}$.[9] When such conditions are too restrictive, then the above definition can easily be relaxed to that of an approximate extremum estimator $\hat{\boldsymbol{\theta}}_{T,S}$ satisfying, for fixed $S \in \mathbb{N}$,

$$Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}) \leq \inf_{\boldsymbol{\theta}\in\Theta_T} Q_{T,S}(\boldsymbol{\theta}) + O_p(\eta_T), \tag{3.3}$$

with $\eta_T \to 0$ as $T \to \infty$. Clearly, setting $O_p(\eta_T) = 0\ \forall T \in \mathbb{N}$ yields an exact sieve extremum estimator. When furthermore the arg min set exists, then the extremum estimator is given by (3.2) above. Now, recall also that a fundamental feature of

and translation invariance which is sometimes unnecessary. More importantly, a norm inducing the product topology on $\mathcal{B}$ might not exist. The fact that $\delta_{\mathcal{B}}$ inherits translation invariance from the norms $\|\cdot\|_{\mathcal{B}_i}\ \forall i$ will be used in the proof of Theorem 3.4.1.

[7] $\mathfrak{B}(\mathcal{B})/\mathfrak{B}(\mathbb{A})$-measurability of maps from a measurable space $(\mathbb{A}, \mathfrak{B}(\mathbb{A}))$ into $(\mathcal{B}, \mathfrak{B}(\mathcal{B}))$ is thus implied by the $\mathfrak{B}(\mathcal{B}_i)/\mathfrak{B}(\mathbb{A})$-measurability of the projections $\pi_i \circ f : \mathbb{A} \to \mathcal{B}_i$ for every $i \in \mathbb{N}$ (see Corollary A.27).

[8] Please note the change in notation. For completeness, the SNPII estimator (and its criterion function) is now indexed by S. In Chapter 2 this notation was avoided for the sake of simplicity and to highlight the generality of the results which applied to sieve extremum estimators in general.

[9] Nothing is gained by letting $Q_{T,S}$ be defined only on the sieves Θ_T, i.e. by letting $Q_{T,S} : \Theta_T \times \Omega \to \mathbb{R}_0^+$, since an agreeing measurable extension is guaranteed to exist on Θ (see e.g. Stinchcombe and White (1992, Lemma 2.14)

the SNPII estimator in (3.2) or (3.3) is its appropriate definition as a minimizer of a divergence defined on the auxiliary parameter space $\mathcal{B}$. In particular, let us define the maps $\hat{\boldsymbol{\beta}}_T : \Omega \to \mathcal{B}$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) : \Omega \to \mathcal{B}$, $\forall \boldsymbol{\theta} \in \Theta$. Each of these consists of a vector of random variables called *auxiliary estimators* or *auxiliary statistics* indexed by $i \in \mathbb{N}$ and taking values on the factor-spaces $\mathcal{B}_i$. The first vector,

$$\hat{\boldsymbol{\beta}}_T := \left(\hat{\boldsymbol{\beta}}^1_T,\ \hat{\boldsymbol{\beta}}^2_T,\ \hat{\boldsymbol{\beta}}^3_T,\ ...\right)$$

collects those estimators $\hat{\boldsymbol{\beta}}^i_T : \Omega \to \mathcal{B}_i$ that are functions of observed data $\mathbf{x_T}$. Auxiliary estimators of interest should be simple to work with in applications and designed so as to possess desirable convergence properties. In particular, they should take values on well chosen (possibly finite dimensional compact) factor spaces $\mathcal{B}_i$ so that they do not suffer from the complications of estimation on large complex spaces. The second vector of auxiliary estimators,

$$\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}) := \left(\frac{1}{S}\sum_{s=1}^{S}\tilde{\boldsymbol{\beta}}^1_{T,s}(\boldsymbol{\theta})\ ,\ \frac{1}{S}\sum_{s=1}^{S}\tilde{\boldsymbol{\beta}}^2_{T,s}(\boldsymbol{\theta})\ ,\ \frac{1}{S}\sum_{s=1}^{S}\tilde{\boldsymbol{\beta}}^3_{T,s}(\boldsymbol{\theta})\ ,\ ...\right)$$

collects (for any given $\boldsymbol{\theta} \in \Theta$) averages of those estimators $\tilde{\boldsymbol{\beta}}^i_T(\boldsymbol{\theta}) : \Omega \to \mathcal{B}_i$ that are functions of the "artificial" sequence of data $\tilde{\mathbf{x}}^s_{\mathbf{T}}(\boldsymbol{\theta})$ drawn from $\mathrm{D}(\boldsymbol{\theta})$. These estimators should have desirable properties $\forall \boldsymbol{\theta} \in \Theta_T$, $T \in \mathbb{N}$. In particular, $\hat{\boldsymbol{\beta}}_T$ should be measurable and converge in suitable manner to a limit point $\boldsymbol{\beta}^*_0 := \left(\boldsymbol{\beta}^*_{0,1},\ \boldsymbol{\beta}^*_{0,2},\ ...\right)$ in $\mathcal{B}$. The random map $\tilde{\boldsymbol{\beta}}_{T,S} : \Omega \times \Theta \to \mathcal{B}$, called *empirical binding function*, should be measurable and converge in an appropriate fashion to a limit deterministic map $\boldsymbol{\beta}^* : \Theta \to \mathcal{B}$, called the *binding function* $\boldsymbol{\beta}^* := \left(\boldsymbol{\beta}^*_1,\ \boldsymbol{\beta}^*_2,\ ...\right)$.[10] Making use of the two vectors of auxiliary estimators $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$, recall that the SNPII estimator's criterion function was defined in Chapter 2 as,

$$Q_{T,S}(\boldsymbol{\theta}) := \mu_T\Big(\hat{\boldsymbol{\beta}}_T\ ,\ \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) \qquad \text{and} \qquad Q_\infty(\boldsymbol{\theta}) = \mu_\infty\Big(\boldsymbol{\beta}^*_0\ ,\ \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big)\ ,\quad \forall\ \boldsymbol{\theta} \in \Theta,$$

where μ_T is a *criterion divergence* and the sequence $\{\mu_T\}_{T\in\mathbb{N}}$ converges in a suitable manner to a *limit criterion divergence* μ_∞. Much notational simplicity in proofs can however be achieved by adopting an alternative more restrictive 'norm-like' criterion divergence μ_T that minimizes the difference $\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$.[11] Hence, here we shall define the SNPII estimator's criterion function $Q_{T,S}$ and its limit Q_∞ as the following real-valued maps,

$$Q_{T,S}(\boldsymbol{\theta}) := \mu_T\Big(\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) \qquad \text{and} \qquad Q_\infty(\boldsymbol{\theta}) = \mu_\infty\Big(\boldsymbol{\beta}^*_0 - \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big)\ ,\quad \forall\ \boldsymbol{\theta} \in \Theta, \tag{3.4}$$

[10] Under Tychonoff's topology on $\mathcal{B}$ this shall be obtained by the appropriate convergence of the projections $\boldsymbol{\beta}^*_i$ in $\mathcal{B}_i$ for every $i = 1, 2, ...$

[11] Note that Assumption 3.1.3 implies that $\mathcal{B}$ is a subset of a vector space. Te difference $\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ should thus be well defined.

where the criterion divergence is now defined as $\mu_T : \mathcal{B} \to \mathbb{R}$ for every $T \in \mathbb{N}$, and its limit as $\mu_\infty : \mathcal{B} \to \mathbb{R}$. Further notational simplification is obtained by defining also the *centered empirical binding function* $\Delta_{T,S}(\boldsymbol{\theta}) := \hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ which can be seen as the natural estimator of the *centered binding function* $\Delta_\infty(\boldsymbol{\theta}) := b(\boldsymbol{\theta}_0) - b(\boldsymbol{\theta})$. Since $\exists \boldsymbol{\theta}_0 \in \Theta : \mathrm{D}(\boldsymbol{\theta}_0) = \mathrm{D}_0$, i.e. $\mathrm{D}_0 \in \mathcal{D}_\Theta$, the $\mathcal{B}$-valued centered binding function $\Delta_\infty : \Theta \to \mathcal{B}$ crosses the origin of $\mathcal{B}$ at $\boldsymbol{\theta}_0$, i.e. $\Delta_\infty(\boldsymbol{\theta}_0) = 0$. Its estimator $\Delta_{T,S} : \Omega \times \Theta \to \mathcal{B}$ does not necessarily cross the origin.

Recall from Chapter 2 that the justification for the use of a sequence of criterion divergences $\{\mu_T\}_{T\in\mathbb{N}}$ instead of a single fixed μ is a practical one. In particular, we note that in applications it is not possible to make use of an infinite number of auxiliary estimators. Hence, μ_T can then be appropriately chosen to be a divergence that gives 'positive weight' only to a finite subset of the vectors $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ for every $T \in \mathbb{N}$, yet converges to a divergence μ_∞ that gives 'positive weight' to the entire vector of auxiliary estimators. For concreteness, let us consider as an example the use of a criterion divergence that extends naturally the classical *indirect inference* setting of Gourieroux et al. (1993). This shall constitute our example of reference and we shall return to it whenever an illustration seems convenient.

Example 3.2.1. *Let μ_T consist of a weighted sum of squared distances between auxiliary estimators that assigns positive weights $w_{T,i}$ to an increasing number k_T of auxiliary estimators,*

$$Q_{T,S}(\boldsymbol{\theta}) = \mu_T\Big(\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) = \sum_{i\in\mathbb{N}} w_{T,i}\Big(\hat{\boldsymbol{\beta}}_T^i - \tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta})\Big)^2 \tag{3.5}$$

where $w_{T,i} > 0$ for every $i \leq k_T$ and $k_T \to \infty$ as $T \to \infty$. Several aspects play a role in ensuring that $Q_{T,S}(\boldsymbol{\theta})$ converges to a well defined (finite) limit $Q_\infty(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \Theta$,

$$Q_\infty(\boldsymbol{\theta}) = \mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) = \sum_{i\in\mathbb{N}} w_i\Big(\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_i^*(\boldsymbol{\theta})\Big)^2. \tag{3.6}$$

Namely, appropriate regularity conditions must be imposed on (i) *the 'speed' at which k_T diverges,* (ii) *the 'decay' of $w_{T,i}$ over i and T, and* (iii) *the stochastic properties of $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$, e.g. the speed of convergence to their limits $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}^*(\boldsymbol{\theta})$.*

Indeed, in the context of Example 3.2.1 above, let $k_T = T$ and $w_i = 1\,\forall i < k_T$. Then $Q_{T,S}(\boldsymbol{\theta})$ will diverge for every $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and $Q_{T,S}(\boldsymbol{\theta}_0)$ will converge only under strict conditions on the convergence rate of $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)$. On the contrary, if $w_{T,i} \to 0$ along T and i in an appropriate fashion, then $Q_{T,S}(\boldsymbol{\theta})$ might remain bounded and converge to a well defined limit $Q_\infty(\boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \Theta$, even under very weak conditions on the stochastic behavior, dependence and heterogeneity of the auxiliary instruments $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)$. These conditions are discussed in the sections below. In particular, Section 3.7 reviews briefly some issues concerned

with obtaining a well-defined Q_∞ by providing appropriate conditions under which weighted sums of converging statistics are themselves well defined and converging to a finite limit. Until then, we simply assume this to be the case. In the context of Example 3.2.1, this can naturally be seen as an implicit restriction on the behavior of k_T, $w_{T,i}$, $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)$.

3.3 Existence and Measurability

As we shall now see, measurability of the SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S} : \Omega \to \Theta_T$ follows almost immediately from *(i)* a measurability result for sieve extremum estimators established by Theorem 2.2 in White and Wooldrige (1991) and *(ii)* the measurability of the auxiliary maps $\hat{\boldsymbol{\beta}}_T : \Omega \to \mathcal{B}$ and $\tilde{\boldsymbol{\beta}}_{T,S} : \Omega \times \Theta \to \mathcal{B}$. The latter requirement is directly obtained from the measurability of the individual auxiliary estimators under the Borel σ-algebra $\mathfrak{B}(\mathcal{B})$ generated by the product topology $\mathcal{T}_{\mathcal{B}}$.

Assumption 3.3.1. (i) $\hat{\boldsymbol{\beta}}_T^i : \Omega \to \mathcal{B}_i$ *is* $\mathcal{F}/\mathfrak{B}(\mathcal{B}_i)$*-measurable* $\forall (T, i) \in \mathbb{N} \times \mathbb{N}$
(ii) $\tilde{\boldsymbol{\beta}}_{T,s}^i(\cdot, \boldsymbol{\theta}) : \Omega \to \mathcal{B}_i$ *is* $\mathcal{F}/\mathfrak{B}(\mathcal{B}_i)$*-measurable* $\forall (\boldsymbol{\theta}, T, s, i) \in \Theta \times \mathbb{N} \times \{1, ..., S\} \times \mathbb{N}$.

We also impose the following continuity assumptions.

Assumption 3.3.2. $\tilde{\boldsymbol{\beta}}_{T,s}^i(\omega, \cdot) : \Theta \to \mathcal{B}_i$ *is continuous on* Θ $\forall (\omega, T, s, i) \in \Omega \times \mathbb{N} \times \{1, ..., S\} \times \mathbb{N}$.

Assumption 3.3.3. $\mu_T : \mathcal{B} \to \mathbb{R}$ *is continuous on* $\mathcal{B}$ $\forall\, T \in \mathbb{N}$.

Theorem 3.3.1 establishes the existence result for the approximate SNPII estimator defined in (3.3).

Theorem 3.3.1. (Existence of SNPII Estimator) *Let Assumptions 3.1.1-3.3.3 hold, then there exists* $\hat{\boldsymbol{\theta}}_{T,S} : \Omega \to \Theta_T$ *satisfying (3.3) and (3.4)* $\forall T \in \mathbb{N}$ *and* $S \in \mathbb{N}$ *that is* $\mathcal{F}/\mathfrak{B}(\Theta_T)$*-measurable.*

The same measurability result applies immediately to the exact SNPII estimator defined in (3.2).

Corollary 3.3.1. (Existence of SNPII Estimator) *Let Assumptions 3.1.1-3.3.3 hold, then there exists a map* $\hat{\boldsymbol{\theta}}_{T,S} : \Omega \to \Theta_T$ *satisfying (3.2) and (3.4)* $\forall T \in \mathbb{N}$ *and* $S \in \mathbb{N}$ *that is* $\mathcal{F}/\mathfrak{B}(\Theta_T)$*-measurable.*

We thus proceed under the established result that $\hat{\boldsymbol{\theta}}_{T,S}$ is a random element taking values in subsets of Θ for every $T \in \mathbb{N}$. Statements involving convergence in law, in probability or almost surely of $\hat{\boldsymbol{\theta}}_{T,S}$ are from now on considered sound under the set of Assumptions 3.1.1-3.3.3.

3.4 Consistency

Recall from Chapter 1 that consistency proofs for the general *sieve extremum estimator* exist under mild regularity conditions that allow for great generality in the choice of sieves and for a variety of forms of dependence and heterogeneity to be present in the data; see e.g. Gallant (1987), White and Wooldrige (1991) and Chen (2007). This section establishes the consistency of the SNPII estimator. In particular, the convergence in probability (and almost surely) of $\hat{\boldsymbol{\theta}}_{T,S}$, as defined in either (3.2) or (3.3), to the parameter $\boldsymbol{\theta}_0 \in \Theta$. As explained in Chapter 2, we can make use of the existing consistency theorems and reduce our task to the verification of its assumptions. Basically, under appropriate regularity conditions, we will proceed to obtain the consistency of the SNPII estimator from *(i)* the uniform convergence of the criterion function $Q_{T,S}$ across the sieves $\Theta_T \; \forall T \in \mathbb{N}$, and *(ii)* the identifiable uniqueness of $\boldsymbol{\theta}_0 \in \Theta$.

Note that in Section 3.1 we have not been precise as to which metric $\delta_{\mathcal{B}}$ is defined on $\mathcal{B}$, requiring only that it induces Tychonoff's topology on the set.[12] One way of obtaining simpler proofs for the theorems that follow, consists of further restricting the class of metrics that are allowed to equip $\mathcal{B}$. In particular, we impose the seemingly mild regularity condition (satisfied e.g. by both metrics in (3.1); see Proposition A.39) that the product metric $\delta_{\mathcal{B}}$ be *Lipschitz weaker* than the *uniform product metric* (see Definitions A.36 and A.38).[13]

Assumption 3.4.1. *$\exists k \in \mathbb{R}^+$ such that $\delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}') \leq k \cdot \sup_{i \in \mathbb{N}} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}'_i\|_{\mathcal{B}_i} \, \forall (\boldsymbol{\beta}, \boldsymbol{\beta}') \in \mathcal{B} \times \mathcal{B}$ where $\boldsymbol{\beta}_i := \pi_i(\boldsymbol{\beta}) \in \mathcal{B}_i$ and $\boldsymbol{\beta}'_i := \pi_i(\boldsymbol{\beta}') \in \mathcal{B}_i$ are projections for every $i \in \mathbb{N}$.*

In what follows we shall make use of some simplifying assumptions. Namely, we assume that the vectors of auxiliary estimators converge to their singleton limits uniformly over $i \in \mathbb{N}$ and across sieves $\{\Theta_T\}_{T \in \mathbb{N}} \subseteq \Theta$. This assumption is useful because it simplifies considerably the proofs and allows us to avoid a number of technical details that can easily distract us from the main consistency argument being conveyed. Nonetheless, it is important to keep in mind the following. First, uniform convergence of auxiliary estimators over $i \in \mathbb{N}$ is not a necessary condition for the consistency results derived here. Second, albeit unnecessarily restrictive, uniform convergence of auxiliary estimators over $i \in \mathbb{N}$ can be easily derived (see Section 3.8 for a discussion of alternative sufficient conditions). Third, uniform convergence of auxiliary estimators over $\boldsymbol{\theta} \in \Theta$ is a typical assumption in *indirect*

[12] Convergence results are nonetheless meaningful. Given metrics $\{\|\cdot\|_{\mathcal{B}_i}\}_{i \in \mathbb{N}}$ on $\{\mathcal{B}_i\}_{i \in \mathbb{N}}$, any pair of product metrics $\delta_{\mathcal{B}}$ inducing Tychonoff's topology on $\mathcal{B}$ is, by definition, topologically equivalent, and convergence in one implies convergence in the other (see Definition A.34 and Remark A.35)).

[13] It is possible that all metrics inducing the product topology satisfy this requirement, in which case Assumption 3.4.1 is redundant. Unfortunately, I find myself unable to prove this statement.

inference. Hence, it does not really carry any new elements. Furthermore, for the reasons covered in Chapter 1, here we actually work under the weaker requirement of uniform convergence across sieves. Finally, uniform convergence jointly over $i \in \mathbb{N}$ and across sieves $\{\Theta_T\}_{T\in\mathbb{N}}$ can also be obtained (see again Section 3.8 for details). Having emphasized the unnecessarily restrictive nature of this assumption, let us now proceed to make deliberate use of it.[14]

Assumption 3.4.2. (i) $\sup_{i\in\mathbb{N}} \left\|\hat{\boldsymbol{\beta}}_T^i - \boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)\right\|_{\mathcal{B}_i} \xrightarrow{p} 0$ *as* $T \to \infty$*;*
(ii) $\sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \left\|\tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta}) - \boldsymbol{\beta}_i^*(\boldsymbol{\theta})\right\|_{\mathcal{B}_i} \xrightarrow{p} 0$ *as* $T \to \infty$ $\forall s \in \{1, ..., S\}$.

In the context of indirect inference, identification of $\boldsymbol{\theta}_0$ requires the fundamental condition that the product binding function $\boldsymbol{\beta}^*$ be injective. This is ensured by having, for every pair $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta \times \Theta$, at least one $i \in \mathbb{N}$ such that the limit $\boldsymbol{\beta}_i^*$ of $\tilde{\boldsymbol{\beta}}_{T,s}^i$ satisfies $\boldsymbol{\beta}_i^*(\boldsymbol{\theta}) \neq \boldsymbol{\beta}_i^*(\boldsymbol{\theta}')$. Furthermore, to ensure the "transfer" of some topological structure from Θ to the factor spaces $\mathcal{B}_i$ (and ultimately to $\mathcal{B}$), we shall assume that the factor binding function $\boldsymbol{\beta}_i^*$ is an open map $\forall i \in \mathbb{N}$. Finally, to guarantee the continuity of the limit criterion function Q_∞ we also impose that $\boldsymbol{\beta}^*$ be continuous on Θ $\forall i \in \mathbb{N}$. Together, these conditions imply that the product binding function $\boldsymbol{\beta}^*$ is a homeomorphism on its range (see proof of Theorem 3.4.1). The parameter space Θ is thus homeomorphic (topologically equivalent) to a subset of $\mathcal{B}$. This conveys a natural sense in which inference on Θ can be conducted through inference on $\mathcal{B}$.

Assumption 3.4.3. $\boldsymbol{\beta}_i^* : \Theta \to \mathcal{B}_i$ *is* (i) *an open map* $\forall i \in \mathbb{N}$*;* (ii) *continuous on* Θ $\forall i \in \mathbb{N}$*; and* (iii) *for every* $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta \times \Theta$, $\exists\, i \in \mathbb{N} : \boldsymbol{\beta}_i^*(\boldsymbol{\theta}) \neq \boldsymbol{\beta}_i^*(\boldsymbol{\theta}')$.

Finally, as we shall see, given Assumption 3.4.3, a sufficient condition for $\boldsymbol{\theta}_0$ to be an identifiably unique minimizer (Definition A.52) of the limit criterion function Q_∞, is that μ_∞ have a well-separated minimum at the origin. In particular, we now require the uniform convergence of the deterministic sequence of criterion divergences $\{\mu_T\}_{T\in\mathbb{N}}$ to a limit criterion divergence μ_∞ that satisfies an identifiable uniqueness condition w.r.t. $0_\mathcal{B} \in \mathcal{B}$ where $0_\mathcal{B}$ denotes the origin of $\mathcal{B}$.

Assumption 3.4.4. *The sequence* $\{\mu_T\}_{T\in\mathbb{N}}$ *satisfies* $\sup_{\boldsymbol{\beta}\in\mathcal{B}} |\mu_T(\boldsymbol{\beta}) - \mu_\infty(\boldsymbol{\beta})| \to 0$ *as* $T \to \infty$ *for some continuous* $\mu_\infty : \mathcal{B} \to \mathbb{R}$.

Assumption 3.4.5. $\inf_{\boldsymbol{\beta}\in S^c(0_\mathcal{B},\epsilon)\subset\mathcal{B}} \left|\mu_\infty(\boldsymbol{\beta}) - \mu_\infty(0_\mathcal{B})\right| > 0 \;\; \forall \epsilon > 0$.

Note here that Assumption 3.4.4 is concerned with the sure convergence of a sequence of well defined deterministic divergences $\{\mu_T\}_{T\in\mathbb{N}}$. This assumption does not address the probabilistic convergence of the random sequence $\{\mu_T(\Delta_{T,S}(\boldsymbol{\theta}))\}_{T\in\mathbb{N}}$

[14] In Assumption 3.4.2 recall that $\boldsymbol{\beta}_i^*(\boldsymbol{\theta}) := \pi_i \circ \boldsymbol{\beta}^*(\boldsymbol{\theta}) \in \mathcal{B}_i \; \forall (\boldsymbol{\theta}, i) \in \Theta \times \mathbb{N}$.

to a limit $\mu_\infty(\Delta_\infty(\boldsymbol{\theta}))$. Also, note that if $\mathcal{B}$ is compact, then the continuity of μ_∞ follows from the uniform convergence of continuous μ_T (recall Assumption 3.3.3). Furthermore, under the compactness of $\mathcal{B}$ and continuity of μ_∞, the identifiable uniqueness of $0_\mathcal{B}$ follows from simple uniqueness (see Chapter 5). For concreteness, let us make use of our reference example with weighted quadratic divergences.

Example 3.4.1. *Let μ_T and μ_∞ be given by (3.5) and (3.6). Then, it is easy to verify that, under appropriate regularity conditions, Assumption 3.4.4 holds for weights satisfying e.g. $w_{T,i} = \mathbb{1}(i < k_T)/2^i$ with $w_i = 1/2^i$ for every $i \in \mathbb{N}$. Furthermore, μ_∞ can be shown to satisfy the identifiable uniqueness condition postulated in Assumption 3.4.5.*

The following theorem provides us with the first asymptotic result of interest.

Theorem 3.4.1. (Consistency) *Let Assumptions 3.1.1-3.4.5 hold. Then, the approximate SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ defined in (3.3) satisfies $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}_0) \xrightarrow{p} 0$ as $T \to \infty$. If Assumption 3.4.2 holds with a.s. convergence, then $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}_0) \xrightarrow{a.s.} 0$.*

The consistency of the exact SNPII estimator follows immediately as a corollary.

Corollary 3.4.1. (Consistency) *Let Assumptions 3.1.1-3.4.5 hold. Then, the exact SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ defined in (3.2) satisfies $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}_0) \xrightarrow{p} 0$ as $T \to \infty$. If Assumption 3.4.2 holds with a.s. convergence, then $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}_0) \xrightarrow{a.s.} 0$.*

Finally, in applications, we might be interested in the study of such quantities as $\left(\phi_1(\boldsymbol{\theta}_0), ..., \phi_{n_\phi}(\boldsymbol{\theta}_0)\right)$ where $\phi_i : \Theta \to \Phi$ is some continuous functional defined on Θ, $i = 1, ..., n_\phi$, $n_\phi \in \mathbb{N}$, be this a map to finite or infinite dimensional spaces. Examples of interest are likely to include finite dimensional objects such as a set of derivatives (when elements $\boldsymbol{\theta}$ are smooth functions) or projections to finite dimensional subsets of Θ (see Andrews (1991) for more examples).

Corollary 3.4.2. *Let the conditions of any of the above Theorem 3.4.1 be satisfied. Let $\hat{\boldsymbol{\theta}}_{T,S}$ denote the corresponding SNPII estimator. Let $\phi : \Theta \to \Phi$ denote a continuous functional (possibly a projection). Then, $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}_0) \xrightarrow{p} 0$ as $T \to \infty$ and by the Continuous Mapping Theorem, $\delta_\Phi(\phi(\hat{\boldsymbol{\theta}}_{T,S}), \phi(\boldsymbol{\theta}_0)) \xrightarrow{p} 0$. Naturally, if $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}_0) \xrightarrow{a.s.} 0$ then $\delta_\Phi(\phi(\hat{\boldsymbol{\theta}}_{T,S}), \phi(\boldsymbol{\theta}_0)) \xrightarrow{a.s.} 0$.*

3.5 Convergence Rate and Asymptotic Normality

As discussed in Chapter 1, results on convergence rates of sieve-estimators are available with some generality for the special case of sieve M-estimators and series estimators. The convergence rate of sieve estimators is typically found to be slow

if the size and/or complexity of the sieves increases slowly with T (in which case approximation errors decrease slowly and dominate) but also, when the entropy of the sieves grows too fast (in which case the estimator's convergence within sieves is typically slow and dominates). Obtaining an appropriate rate of convergence of sieve estimators thus usually requires a balance between the approximation error and the rate of convergence of the estimator within each sieve. This section establishes the $\sqrt{T}$-convergence rate and asymptotic normality for the SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$. As revealed by Chapter 2, greater generality could be achieved by providing separate conditions for the convergence rate and the asymptotic distribution. We do not pursue this here for the sake of brevity and simplicity.

Remark 3.5.1. *In essence, the $\sqrt{T}$-convergence rate and asymptotic normality will be derived from an equivalent asymptotic behavior of the individual auxiliary estimators $\hat{\boldsymbol{\beta}}^i_T$ and $\tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta}_0)$ for every $i \in \mathbb{N}$.*

To achieve the desired results, a number of appropriate regularity conditions must however be added. First of all, Θ is now required to be a normed vector space (hence a separable Banach space taking into account Assumption 3.1.1). The linear space structure is required for us to make use of differentiability and linearity concepts that are an integral part of the theory that follows. It is on linear spaces that such maps are naturally defined. Second, the auxiliary statistics are assumed to be random variables taking values in $\mathbb{R}^{q_i}$. Accordingly, auxiliary factor spaces $\mathcal{B}_i$ are assumed to be compact subsets of $\mathbb{R}^{q_i}$ $\forall i \in \mathbb{N}$. Compactness not only simplifies proofs, it enables the use of several well-established results of weak convergence on compact sets.

Assumption 3.5.1. $(\Theta, \|\cdot\|_\Theta)$ *is a subset of a normed vector space.*[15]

Assumption 3.5.2. $(\mathcal{B}_i, \|\cdot\|_{\mathcal{B}_i})$ *is a compact subset of* $\mathbb{R}^{q_i}$ $\forall i \in \mathbb{N}$.[16]

To derive the $\sqrt{T}$-convergence rate and asymptotic normality of $\hat{\boldsymbol{\theta}}_{T,S}$ we shall make use of a number of smoothness conditions which turn out to be sufficient to obtain a Z-estimator formulation of our extremum estimator $\hat{\boldsymbol{\theta}}_{T,S}$. In particular, $\hat{\boldsymbol{\theta}}_{T,S}$ can be shown to set all first order directional derivatives of the criterion function $Q_{T,S}$ to zero (at least approximately). For concreteness, let us follow Chapter 2 and define $\nabla Q_{T,S}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ and $\nabla Q_\infty(\boldsymbol{\theta}, \boldsymbol{\theta}')$ as directional derivatives of $Q_{T,S}$ and Q_∞ at $\boldsymbol{\theta}$

[15] δ_Θ in Assumption 3.1.1 is thus the metric induced by $\|\cdot\|_\Theta$ according to $\delta_\Theta(\boldsymbol{\theta}, \boldsymbol{\theta}') := \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\Theta \; \forall (\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta \times \Theta$. The denseness of the sieves postulated in Assumption 3.1.2 also w.r.t. $\|\cdot\|_\Theta$.

[16] Recall that existence of a norm generating the product topology on a infinite product space is put on doubt by the fact that no Banach space has the Heine-Borel property but that on a countable product space with product topology the unit cube is naturally compact as a product of compact sets. Each of the $\mathcal{B}_i$'s may hence be normed but not necessarily $\mathcal{B}$.

in the direction of $\boldsymbol{\theta}'$. We have seen that, under appropriate regularity conditions, a Z-estimator formulation of $\hat{\boldsymbol{\theta}}_{T,S}$ can be obtained, in the sense that $\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}') = o_p(T^{-1/2})$ and $\nabla Q_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}') = 0$ for every $\boldsymbol{\theta}' \in \text{lin}(\Theta)$.

Sometimes, setting only the partial derivatives (i.e. derivatives in the direction of basis vectors) to zero is sufficient (Lemma C.10). One issue that arises naturally in the present case of an infinite dimensional parameter space Θ is hence that of the existence of a basis for Θ. Being infinite dimensional, the standard notion of a *Hamel* basis is simply unavailable. Nonetheless, Θ might still possess a *Schauder* basis (Definition A.86), in which case we can still reduce our attention to a countable set of derivatives (those in the direction of the *Schauder* basis vectors of Θ) playing the statistical role of an infinite system of estimating equations as in van der Vaart (1995) and van der Vaart and Wellner (1996, ch. 3.3).

Remark 3.5.2. *Most spaces of interest admit a Schauder basis, e.g.* $C([0,1], \sup)$, L^p *and* l^p, $1 \leq p < \infty$ *and every separable Hilbert space with an orthonormal basis. However, existence of a Schauder basis for* Θ *cannot be guaranteed from the outset under the current set of assumptions.*[17]

From a theoretical point of view, the existence of the Schauder basis is not strictly necessary. We shall nonetheless impose it so as to obtain simpler results, make them intuitively more appealing, and also, to avoid a number of problems of a more philosophical nature. With this in mind, we let $\mathbb{S}_\Theta$ denote the Schauder basis of Θ. A Z-estimator formulation of the SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ is thus available where $\boldsymbol{\theta}_0$ is a root of $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta) : \Theta \to \mathbb{R}^{|\mathbb{S}_\Theta|}$.[18] The infinite dimensionality of Θ ensures that $\mathbb{R}^{|\mathbb{S}_\Theta|} \equiv \mathbb{R}^\infty$. The infinite system of estimating equations satisfies naturally $\nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta) = 0_{\mathbb{S}_\Theta}$ where $0_{\mathbb{S}_\Theta}$ denotes the zero of the *system's image set* $\mathbb{R}^{|\mathbb{S}_\Theta|}$.

Just as in the finite-dimensional case, the differentiability of the infinite system of partial derivatives shall be derived from the differentiability of each partial derivative individually, by letting the *system's image set* $\mathbb{R}^{|\mathbb{S}_\Theta|}$ be equipped with the *product topology*. Estimation of $\boldsymbol{\theta}_0$ can thus be described by finding roots $\hat{\boldsymbol{\theta}}_{T,S}$ of $\nabla Q_{T,S}(\cdot, \mathbb{S}_{\Theta_T}) : \Theta \to \mathbb{R}^{|\mathbb{S}_{\Theta_T}|}$ for every $T \in \mathbb{N}$. Finite dimensional sieves Θ_T possess a finite Schauder basis $\mathbb{S}_{\Theta_T}$ that coincides with the usual Hamel basis, i.e. $|\mathbb{S}_{\Theta_T}| < \infty$. Estimation is thus further characterized by the use of an increasing system of random estimating equations with $\hat{\boldsymbol{\theta}}_{T,S}$ satisfying $\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_{\Theta_T}) = o_p(T^{-1/2})$. See Remark C.2 in Appendix C for details on the construction of $\nabla Q_{T,S}(\cdot, \mathbb{S}_{\Theta_T})$.

Assumption 3.5.3. $(\Theta, \|\cdot\|_\Theta)$ *admits a Schauder basis* $\mathbb{S}_\Theta \subseteq \Theta$. *The image set* $\mathbb{R}^{|\mathbb{S}_\Theta|}$ *is endowed with Tychonoff's topology* $\mathcal{T}_{\mathbb{R}^{|\mathbb{S}_\Theta|}}$.

[17] Recall that even a separable Banach space might fail to possess a Schauder basis. A theorem of Mazur asserts however that every infinite-dimensional Banach space has an infinite-dimensional subspace with a Schauder basis.

[18] Following van der Vaart and Wellner (1996), Q_∞ is allowed to have multiple roots.

The complicated convergence structure of sieve estimators does not always help in keeping with clarity and intuition. Some light can however be shed on this issue by decomposing, for every $(\omega, T, S) \in \Omega \times \mathbb{N} \times \mathbb{N}$, the distance $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\|_\Theta$ into parts that explicitly identify the minimizers of the criterion function $Q_{T,S}$ and its limit Q_∞ on both Θ_T and Θ. In particular, for every $\omega \in \Omega$, let $\boldsymbol{\theta}^*_{T,S}$ and $\boldsymbol{\theta}^{**}_{T,S}$ denote elements of the arg min set of $Q_{T,S}$ over the sieve Θ_T and the entire parameter space Θ respectively,

$$\boldsymbol{\theta}^*_{T,S} \in \arg\min_{\boldsymbol{\theta}\in\Theta_T} Q_{T,S}(\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\theta}^{**}_{T,S} \in \arg\min_{\boldsymbol{\theta}\in\Theta} Q_{T,S}(\boldsymbol{\theta}) \;\; \forall (T,S) \in \mathbb{N}\times\mathbb{N}.$$

Furthermore, recall that $\boldsymbol{\theta}_0$ is the unique minimizer of Q_∞ over Θ, and let $\boldsymbol{\theta}^0_T$ denote an element of the arg min set of Q_∞ over Θ_T for every $T \in \mathbb{N}$,

$$\boldsymbol{\theta}^0_T \in \arg\min_{\boldsymbol{\theta}\in\Theta_T} Q_\infty(\boldsymbol{\theta}) \;\; \forall T \in \mathbb{N}.$$

Then, for every $(T,S) \in \mathbb{N}\times\mathbb{N}$ it holds surely ($\forall \omega \in \Omega$) that,

$$\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\|_\Theta \;\leq\; \|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta + \|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta + \|\boldsymbol{\theta}^{**}_{T,S} - \boldsymbol{\theta}^0_T\|_\Theta + \|\boldsymbol{\theta}^0_T - \boldsymbol{\theta}_0\|_\Theta. \tag{3.7}$$

As noted in Chapter 2, this decomposition turns out to be especially useful as it allows us to separate the study of an upper bound on the convergence rate of $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\|_\Theta$ into *(i)* the convergence rate of the "approximation error" $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta$ introduced by having $\hat{\boldsymbol{\theta}}_{T,S}$ be an *approximate* extremum estimator; *(ii)* the convergence rate of the "error" $\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta$ introduced by constraining the optimization of $Q_{T,S}$ to the sieves Θ_T; *(iii)* the distance $\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^0_T\|_\Theta$ between the global minimizer of $\boldsymbol{\theta}^*_{T,S}$ of $Q_{T,S}$ and the constrained minimizer $\boldsymbol{\theta}^0_T$ of its limit Q_∞ ; and *(iv)* the distance $\|\boldsymbol{\theta}^0_T - \boldsymbol{\theta}_0\|_\Theta$ between the restricted and unrestricted minimizer of Q_∞.

It should now be evident from (2.19) that, depending on the exact nature of the SNPII estimator, different conditions will be required to derive an appropriate convergence rate for $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\|_\Theta$. For example, exact SNPII estimators (as defined in (3.2)) satisfy naturally $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta = 0$. Also, SNPII estimators taking values on *purely dimensional* sieves w.r.t. $\{Q_{T,S}\}_{T\in\mathbb{N}}$ (see Definition A.82) satisfy $\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta = 0$.[19] In what follows we derive separate convergence rate theorems depending on the nature of the SNPII estimator. First, however, we derive some preliminary results that are common to these alternative formulations of the estimator.

3.5.1 Some Common Preliminary Results

The first smoothness requirements of interest to us are those of *continuous Hadamard differentiability* (CHD) and *uniform Hadamard equi-differentiability of the third kind*

[19] This is likely to be the most common formulation of SNPII estimation problems in applications.

(UHED3) (see Definitions C.3, C.18, C.23 and C.24). Assumption 3.5.4 below establishes primitive smoothness conditions. Items *(i)* and *(ii)* impose smoothness conditions on auxiliary estimators $\tilde{\boldsymbol{\beta}}^{i}_{T,S}$ and respective derivatives $\nabla\tilde{\boldsymbol{\beta}}^{i}_{T,S}$. Items *(iii)* and *(iv)* establish smoothness conditions on the binding function $\boldsymbol{\beta}^{*}_{i}$ and its derivative $\nabla\boldsymbol{\beta}^{*}_{i}$. Items *(v)* and *(vi)* focus on the criterion divergence μ_T and respective derivative $\nabla\mu_T$. Finally, items *(vii)* and *(viii)* provide smoothness conditions on the limits μ_∞ and $\nabla\mu_\infty$.

Assumption 3.5.4. (i) $\tilde{\boldsymbol{\beta}}^{i}_{T,S} : \Omega \times \Theta \to \mathcal{B}_i$ *is a.s. CHD on* $\Theta\,\forall (T,i) \in \mathbb{N}\times\mathbb{N}$ *and* $\{\tilde{\boldsymbol{\beta}}^{i}_{T,S}\}_{T\in\mathbb{N}}$ *is a.s. UHED3 along sequences* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0 \;\; \forall i \in \mathbb{N}$*;* (ii) *for every* $\boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta$ $\{\nabla\tilde{\boldsymbol{\beta}}^{i}_{T,S}(\cdot,\boldsymbol{\theta}'_T)\}_{T\in\mathbb{N}}$ *is a.s. UHED3 along sequences* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0 \;\; \forall i \in \mathbb{N}$*;* (iii) $\boldsymbol{\beta}^{*}_{i} : \Theta \to \mathcal{B}_i$ *is CHD on* $\Theta\,\forall i \in \mathbb{N}$*;* (iv) $\nabla\boldsymbol{\beta}^{*}_{i}(\cdot,\boldsymbol{\theta}) : \Theta \to \mathcal{B}_i$ *is CHD on* $S(\boldsymbol{\theta}_0,\epsilon)$*,* $\epsilon > 0$*,* $\forall(\boldsymbol{\theta},i) \in \text{lin}\Theta\times\mathbb{N}$ *and for every* $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}' \in \Theta$*, the sequence* $\{\nabla\boldsymbol{\beta}^{*}_{i}(\cdot,\boldsymbol{\theta}'_T)\}_{T\in\mathbb{N}}$ *is UHED3 along sequences* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0 \;\; \forall i \in \mathbb{N}$*;* (v) $\mu_T : \mathcal{B} \to \mathbb{R}$ *is CHD on* $\mathcal{B}\,\forall T \in \mathbb{N}$*;* (vi) $\nabla\mu_T : \mathcal{B}\times\mathcal{B} \to \mathbb{R}$ *is CHD on* $\mathcal{B}\times\mathcal{B}\,\forall T \in \mathbb{N}$ *and* $\{\mu_T\}_{T\in\mathbb{N}}$ *is UHED3 along sequences* $\boldsymbol{\beta}^{\nabla}_{T} \to (\Delta_\infty(\boldsymbol{\theta}_0), \nabla\Delta_\infty(\boldsymbol{\theta}_0,\boldsymbol{\theta})) \in \mathcal{B}\times\mathcal{B}$*,* $\boldsymbol{\theta} \in \Theta$*;* (vii) $\mu_\infty : \mathcal{B} \to \mathbb{R}$ *is CHD on* $\mathcal{B}$ *and* (viii) $\nabla\mu_\infty : \mathcal{B}\times\mathcal{B} \to \mathbb{R}$ *is CHD on* $\mathcal{B}\times\mathcal{B}$*.*

Assumptions 3.1.1-3.1.3 and 3.5.1-3.5.4 allow us to obtain the first preliminary result of interest.

Proposition 3.5.1. (Criterion Function Differentiability) *Let Assumptions 3.1.1-3.1.3 and 3.5.1-3.5.4 hold. Then,* (i) $Q_{T,S} : \Omega\times\Theta \to \mathbb{R}$ *is a.s. CHD on* $\Theta\,\forall(T,S) \in \mathbb{N}\times\mathbb{N}$*;* (ii) *for every* $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}$*, the sequence* $\{\nabla Q_{T,S}(\cdot,\boldsymbol{\theta}'_T)\}_{T\in\mathbb{N}}$ *is a.s. UHED3 along sequences* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$*;* (iii) $\{\nabla Q_{T,S}(\cdot,\mathbb{S}_{\Theta_T})\}_{T\in\mathbb{N}}$ *is a.s. UHED3 along sequences* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$*;* (iv) $Q_\infty : \Theta \to \mathbb{R}$ *is CHD on* Θ*;* (v) $\nabla Q_\infty(\cdot,\mathbb{S}_\Theta) : \Theta \to \mathbb{R}^{|\mathbb{S}_\Theta|}$ *is continuously Hadamard differentiable on* $S(\boldsymbol{\theta}_0,\epsilon)$ *for some* $\epsilon > 0$*.* (vi) *for every* $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}$*, the sequence* $\{\nabla Q_\infty(\cdot,\boldsymbol{\theta}'_T)\}_{T\in\mathbb{N}}$ *is UHED3 along sequences* $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$*.*

A second result of interest is concerned with *continuous invertibility* of the appropriately defined second-order Hadamard derivative of the limit criterion function Q_∞. In what follows, we make use of primitive conditions involving *continuous invertibility* of operators (Definition B.12).

Assumption 3.5.5. (i) $\boldsymbol{\beta}^{*}_{i} : \Theta \to \mathcal{B}$ *has a continuously invertible Hadamard derivative at* $\boldsymbol{\theta}_0$ *for every* $i \in \mathbb{N}$*;* (ii) $\nabla\boldsymbol{\beta}^{*}_{i}(\cdot,\boldsymbol{\theta}) : \Theta \to \mathcal{B}$ *has a continuously invertible Hadamard derivative at* $\boldsymbol{\theta}_0$ *for every* $(i,\boldsymbol{\theta}) \in \mathbb{N}\times\text{lin}\Theta$*;* (iii) $\nabla\mu_\infty : \mathcal{B}\times\mathcal{B} \to \mathbb{R}$ *has a continuously invertible Hadamard derivative at* $\Big(\Delta_\infty(\boldsymbol{\theta}_0), \nabla\Delta_\infty(\boldsymbol{\theta}_0,\boldsymbol{\theta})\Big) \in \mathcal{B}\times\mathcal{B}$ *for every* $\boldsymbol{\theta} \in \Theta$*.*

The following Proposition establishes a desirable implication of Assumption 3.5.5. Namely, the *continuous invertibility* of $\nabla Q^{\nabla_{\boldsymbol{\theta}}}_{\infty}(\boldsymbol{\theta}_0,\cdot) \in \mathbb{L}(\text{lin}\Theta,\mathbb{R})$ for every $\boldsymbol{\theta} \in \text{lin}\Theta$ where $Q^{\nabla_{\boldsymbol{\theta}}}_{\infty} := \nabla Q_\infty(\cdot,\boldsymbol{\theta}) : \Theta \to \mathbb{R}$. A second result of interest follows immediately as a corollary by appealing to Proposition B.16. Namely

that $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathbb{R}^{|\mathbb{S}_\Theta|})$ is also *continuously invertible*, where $Q_\infty^{\nabla_{\mathbb{S}_\Theta}} := \nabla Q_\infty(\cdot, \mathbb{S}_\Theta) : \Theta \to \mathbb{R}^{|\mathbb{S}_\Theta|}$.

Proposition 3.5.2. (Continuous Invertibility of Limit Criteria) *Let Assumptions 3.1.1-3.1.3, 3.5.1-3.5.3 and 3.5.5 hold. Then, $\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathbb{R})$ is continuously invertible $\forall \boldsymbol{\theta} \in \text{lin}\Theta$. Furthermore, $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathbb{R}^{|\mathbb{S}_\Theta|})$ is also continuous invertible.*

Assumption 3.5.4 reveals an important cost of introducing sieves. Namely, while smoothness requirements are typically required to hold only at $\boldsymbol{\theta}_0$. Here, the stronger uniform and equi-differentiability concepts must hold. Assumption 3.5.5 on the other hand is quite common. These invertibility requirements are essentially related to e.g. non-singularity of the score in MLE. In general, it is easy to select criterion divergences and auxiliary estimators satisfying such conditions. To fix ideas, it might be interesting to return to our reference example.

Example 3.5.1. *Under suitable mild regularity conditions, the weighted quadratic criterion divergence μ_T in (3.5) and its limit μ_∞ in (3.6) can be shown to satisfy the differentiability conditions of Assumption 3.5.4; see Sundaresan (1967) and Leonard and Sundaresan (1974). Invertibility of the second derivative of μ_∞ is trivially satisfied. The smoothness and invertibility conditions on the auxiliary estimators are obtained from the study for their influence function. Let $\tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\beta} \in \mathcal{B}_i} \sum_{t=1}^T \mathcal{Q}^i\left(\tilde{\mathbf{x}}_{\mathbf{T}}^s(\boldsymbol{\theta}), \boldsymbol{\beta}\right)$ be a real-valued M-estimator. The classic robust-statistics textbook of Huber (1974) reveals that the influence function of the M-estimator $\tilde{\boldsymbol{\beta}}_{T,s}^i$ is proportional to $\mathcal{Q}^i : \mathcal{B}_i \to \mathbb{R}$. The desired result is thus obtained under appropriate conditions on both $\mathcal{Q}^i$ and the distribution of $\tilde{\mathbf{x}}_{\mathbf{T}}^s$.*

The differentiability and invertibility results obtained above allow us to derive an appropriate convergence rate for the term $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta$ in (3.7).[20] In applications, the choice of sieves $\{\Theta_T\}_{T \in \mathbb{N}}$ and norm $\|\cdot\|_\Theta$ is typically guided by the existence of results stemming from the field of *Approximation Theory* establishing *(i)* the denseness of the sequence of sieves $\{\Theta_T\}_{T \in \mathbb{N}}$ on Θ w.r.t. $\|\cdot\|_\Theta$ and *(ii)* the convergence rate of the sequence of projections $\|\pi_{\Theta_T}(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|_\Theta$ as $T \to \infty$.[21] The appropriate convergence

[20] Note here that, by definition, each $\boldsymbol{\theta}_T^0$ corresponds to an element of the projection set of $\boldsymbol{\theta}_0$ onto Θ_T w.r.t. the divergence Q_∞. That Q_∞ is a divergence on Θ w.r.t. $\boldsymbol{\theta}_0$ follows from its definition in terms of the divergence μ_∞ and the injective nature of the product binding function $\boldsymbol{\beta}^*$ derived in Proposition A.46 from Assumption 3.4.3. Existence of the arg min set follows immediately, by Weierstrass's *Extreme Value Theorem* (Lemma A.85), from the compactness of each Θ_T (Assumption 3.1.2) and continuity of Q_∞. Uniqueness of $\boldsymbol{\theta}_T^0$ (i.e. reduction of the arg min set to a singleton) follows for norm-divergences μ_∞ by the strict convexity of μ (or μ_∞ respectively); see Chapter 5.

[21] No similar results are however necessarily available for the sequence $\{\boldsymbol{\theta}_T^0\}_{T \in \mathbb{N}}$ of projections w.r.t. the (probably very complex) divergence Q_∞.

rate of $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta$ is obtained by imposing a primitive "rate of expansion" of the sieves, as described by the convergence rate of $\|\pi_{\Theta_T}(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|_\Theta$.

Assumption 3.5.6. $\|\pi_{\Theta_T}(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|_\Theta = o(T^{-1/2})$ *as* $T \to \infty$.

Proposition 3.5.3. (Convergence Rate of Limit Divergence Projections) *Let Assumptions 3.1.1-3.1.3 and 3.5.1-3.5.6 hold. Then we obtain* $\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta = o(T^{-1/2})$ *and* $\nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) = o\Big(T^{-1/2}\Big)$ *as* $T \to \infty$.

Example 3.5.2. *For simplicity, suppose that our object of interest* $\boldsymbol{\theta}_0$ *consists of a single real-valued function defined on* $[0,1]$. *Let* $\Theta \equiv \mathcal{H}^p([0,1])$ *the Hölder space of* p*-smooth functions and* Θ_T *be spanned by the power monomials up to power* k_T *for every* $T \in \mathbb{N}$. *Then* $\|\pi_{\Theta_T}(\boldsymbol{\theta}_0) - \boldsymbol{\theta}_0\|_\Theta = O(k_T^{-p})$ *for every* L_{p^*}*-norm* $\|\cdot\|_\Theta$ *(for any* p^**). As such, if e.g.* $\boldsymbol{\theta}_0 \in H^3([0,1])$, *we are forced by Assumption 3.5.6 to adopt sieves* Θ_T *spanned by power monomials of order* k_T *growing faster than* $O(T^{1/6})$.[22] *For a* $\boldsymbol{\theta}_0$ *on the smaller space* $\Theta \subseteq H^5([0,1])$, *Assumption 3.5.6 estimation requires a very slow* k_T *only faster than* $O(T^{1/10})$. *On the larger parameter space* $\boldsymbol{\theta}_0 \in H^2([0,1])$ *the requirement is more stringent and possibly troublesome with* k_T *faster than* $O(T^{1/4})$. *This example reveals neatly the price to pay for generality. Details on rates of convergence for alternative parameter spaces and sieves can be found e.g. in Powell (1981), Judd (1998) and Chen (2007).*

A forth preliminary result of crucial importance concerns the weak convergence of the appropriately standardized set of criterion derivatives to a well-defined tight limit Gaussian process $\mathbb{G}_S(\boldsymbol{\theta}_0)$,[23]

$$\sqrt{T}\Big[\nabla Q_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\Big] \xrightarrow{d} \mathbb{G}_S(\boldsymbol{\theta}_0) \quad \text{as} \quad T \to \infty. \tag{3.8}$$

A requirement similar to that in (3.8) is typical also in finite-dimensional M-estimators. Indeed, (3.8) corresponds e.g. to the well known asymptotic normality of the score in finite dimensional ML estimators. As pointed out earlier, $\nabla_{\Theta_T} Q_{T,S}(\cdot, \mathbb{S}_{\Theta_T}) : \Theta \to \mathbb{R}^{|\mathbb{S}_{\Theta_T}|}$ has the interpretation of a (finite) system of real-valued estimating equations with $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta) : \Theta \to \mathbb{R}^{|\mathbb{S}_\Theta|}$ corresponding to the limit infinite system satisfying $\nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta) = 0_{\mathbb{S}_\Theta}$. Proposition C.34 shows that (3.8) can be obtained from the asymptotic normality of $\sqrt{T}\Big(\Delta_{T,S}(\boldsymbol{\theta}_0) - \Delta_\infty(\boldsymbol{\theta}_0)\Big)$ by ensuring that $\Big\{\nabla\mu_T(\cdot, \nabla\Delta_{T,S}(\boldsymbol{\theta}_0), \mathbb{S}_{\Theta_T})\Big\}_{T\in\mathbb{N}}$ is a $\sqrt{T}$*-Hadamard sequence*; see Definition C.21 and Remark 3.5.3 below.

Remark 3.5.3. *Note here that (3.8) is obtained from Proposition C.34 by setting* $(1/t_n) = \sqrt{T}$, $f_T = \nabla\mu_T(\cdot, \nabla\Delta_{T,S}(\boldsymbol{\theta}_0), \mathbb{S}_{\Theta_T})\ \forall T \in \mathbb{N}$ *and* $f = \nabla\mu_\infty(\cdot, \nabla\Delta_\infty(\boldsymbol{\theta}_0), \mathbb{S}_\Theta)$ *with* $X_n - a_\nabla = \Delta_{T,S}(\boldsymbol{\theta}_0) - \Delta_\infty(\boldsymbol{\theta}_0)$ *and* $Z = \mathbb{G}_\Delta$ *a Gaussian process.*

[22] Faster that $O(T^{1/6})$ meaning that $k_T^{-1} = o(T^{-1/6})$.

[23] The index S underlines that the variance of the limit process is typically dependent on S.

A characterization of $\sqrt{T}$-*Hadamard sequences* is provided in Section C.3.2 of Appendix C. The following assumption is thus shown to establish a possible set of sufficient conditions for (3.8).

Assumption 3.5.7. *For fixed* $S \in \mathbb{N}$,
(i) $\sqrt{T}(\hat{\boldsymbol{\beta}}_T^i - \boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma_0^i) \;\; \forall i \in \mathbb{N}$ *as* $T \to \infty$*;*
(ii) $\sqrt{T}(\tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma_S^i(\boldsymbol{\theta}_0)) \;\; \forall i \in \mathbb{N}$ *as* $T \to \infty$*;*
(iii) $\|\nabla \tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla \boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\|_{\mathcal{B}_i^{|\mathbb{S}_\Theta|}} = o_p(r_T) \; \forall i \in \mathbb{N}$ *as* $T \to \infty$*;*
(iv) $\sup_{(\boldsymbol{\beta},\boldsymbol{\beta}') \in \mathcal{B}\times\mathcal{B}} \left|\nabla \mu_T(\boldsymbol{\beta}, \boldsymbol{\beta}') - \nabla \mu_\infty(\boldsymbol{\beta}, \boldsymbol{\beta}')\right| = o(T^{-1/2})$ *as* $T \to \infty$*;*
(v) $\sup_{\boldsymbol{\beta} \in \mathcal{B}} \left|\nabla \mu_\infty(\boldsymbol{\beta}, \boldsymbol{\beta}_T)\right| = o\big(\xi_\mu(\|\boldsymbol{\beta}_T\|_{\mathcal{B}})\big)$ *for every* $\boldsymbol{\beta}_T \to 0$ *and* $r_T = \xi_\mu^{-1}(T^{-1/2})$.

Note that in the context of our reference Example 3.2.1, conditions *(iii)* and *(iv)* of Assumption 3.5.7 above impose strict rates of divergence for $k_T \to \infty$ and of the sieve's expansion in terms of dimensionality.

Indeed, it is easy to show that $\left\|\nabla \tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla \boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\right\| = o_p(r_T)$, is (in general) related to the satisfaction of a minimum sieve expansion rate that controls the speed at which the number dimension of Θ_T grows as $T \to \infty$. In particular, this condition can be substituted by simpler counterparts involving a minimum convergence speed for the partial derivatives, a bound on the size of the derivatives of the binding function and a growth speed of the sieve's dimension $p_T := \dim(\Theta_T)$.

Similarly, $\sup_{(\boldsymbol{\beta},\boldsymbol{\beta}') \in \mathcal{B}\times\mathcal{B}} \left|\nabla \mu_T(\boldsymbol{\beta}, \boldsymbol{\beta}') - \nabla \mu_\infty(\boldsymbol{\beta}, \boldsymbol{\beta}')\right| = o(T^{-1/2})$ is (in general) related to the speed at which new auxiliary estimators are given positive weight by the sequence of μ_T's (which is controlled by the variable k_T in Example 3.2.1).

Proposition 3.5.4. (Asymptotic Gaussianity of Criterion Derivative) *Let Assumptions 3.1.1-3.4.5 and 3.5.1-3.5.7 hold. Then, for fixed* $S \in \mathbb{N}$*, (3.8) holds true as* $T \to \infty$ *where* $\mathbb{G}_S(\boldsymbol{\theta}_0)$ *a tight Gaussian process.*

At this point, it is important to note that the same result might be obtained under considerably weaker conditions depending when further properties of $\nabla \mu_T$ and $\nabla \mu_\infty$ are known. This is indeed the case if we refer back to our reference example.

Example 3.5.3. *Consider again our reference example in (3.5) and (3.6). There, both* $\nabla \mu_T$ *and* $\nabla \mu_\infty$ *are bilinear. By Proposition C.35, conditions* (iii) *and* (iv) *in Assumption 3.5.7 can then be substituted by the weaker alternative conditions* $\|\nabla_{\Theta_T} \tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla_\Theta \boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\|_{\mathcal{B}_i} = o_p(1) \; \forall (S, i) \in \mathbb{N} \times \mathbb{N}$ *and*

$$\sup_{(\boldsymbol{\beta},\boldsymbol{\beta}') \in \mathcal{B}\times\mathcal{B}} \left|\nabla \mu_T(\boldsymbol{\beta}, \boldsymbol{\beta}') - \nabla \mu_\infty(\boldsymbol{\beta}, \boldsymbol{\beta}')\right| = o(1)$$

respectively. This example is interesting in that it mimics closely the requirements in Gourieroux et al. (1993). Note also that this considerable simplification can be well understood in the context of Hadamard sequences *by appealing to Propositions C.3.1 and C.35.*

Finally, let us move to the last preliminary result of interest. Namely that,

$$\sqrt{T}\left\|\left(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}-Q_{\infty}^{\nabla_{\mathbb{S}_{\Theta}}}\right)(\hat{\boldsymbol{\theta}}_{T,S})-\left(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}-Q_{\infty}^{\nabla_{\mathbb{S}_{\Theta}}}\right)(\boldsymbol{\theta}_T^0)\right\|=o_p\big(1+\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_{\Theta}\big). \tag{3.9}$$

Alternative sets of conditions can be devised to ensure that (3.9) holds true. Proposition 3.5.5 below makes use conditions on the convergence of second-order derivatives of auxiliary estimators and the criterion divergence. In what follows we let *(i)* $\tilde{\boldsymbol{\beta}}_{T,S}^{\nabla_{\boldsymbol{\theta}},i}(\boldsymbol{\theta}') := \nabla\tilde{\boldsymbol{\beta}}_{T,S}^{i}(\boldsymbol{\theta}',\boldsymbol{\theta})$ for every $(\boldsymbol{\theta}',\boldsymbol{\theta})\in\Theta\times\text{lin}(\Theta)$; *(ii)* $\boldsymbol{\beta}_{\nabla_{\boldsymbol{\theta}}}^{*,i}(\boldsymbol{\theta}') := \nabla\boldsymbol{\beta}_{i}^{*}(\boldsymbol{\theta}',\boldsymbol{\theta})$; *(iii)* $\mu_T^{\nabla}(\boldsymbol{\beta},\boldsymbol{\beta}') := \nabla\mu_T(\boldsymbol{\beta},\boldsymbol{\beta}')$ and *(iv)* $\mu_{\infty}^{\nabla}(\boldsymbol{\beta},\boldsymbol{\beta}') := \nabla\mu_{\infty}(\boldsymbol{\beta},\boldsymbol{\beta}')$ for every $(\boldsymbol{\beta},\boldsymbol{\beta}')\in\mathcal{B}\times\text{lin}(\mathcal{B})$.

Assumption 3.5.8. *Let* (i) $\|\nabla\tilde{\boldsymbol{\beta}}_{T,S}^{\nabla_{\boldsymbol{\theta}},i}(\boldsymbol{\theta}_0,\boldsymbol{\theta}')-\nabla\boldsymbol{\beta}_{\nabla_{\boldsymbol{\theta}}}^{*,i}(\boldsymbol{\theta}_0,\boldsymbol{\theta}')\|=o_p(1)$ *as* $T\to\infty$ *for every* $(\boldsymbol{\theta},\boldsymbol{\theta}',i)\in\mathbb{S}_{\Theta}\times\text{lin}(\Theta)\times\mathbb{N}$*; and* (ii) $\sup_{\boldsymbol{\beta}_{\nabla}\in\mathcal{B}_{\nabla}^{*}}|\nabla\mu_T^{\nabla}(\boldsymbol{\beta}_{\nabla})-\nabla\mu_{\infty}^{\nabla}(\boldsymbol{\beta}_{\nabla})|=o(1)$ *as* $T\to\infty$ *for every compact* $\mathcal{B}_{\nabla}^{*}\subset\text{lin}(\mathcal{B}\times\text{lin}(\mathcal{B})\times\text{lin}(\mathcal{B})\times\text{lin}(\mathcal{B}))$.

Proposition 3.5.5. (Negligible Remainder on Taylor Expansion) *Let Assumptions 3.1.1-3.4.5, 3.5.1-3.5.5 and 3.5.8 hold true. Then the negligible remainder condition in (3.9) holds true.*

3.5.2 Weak Convergence

The following theorem establishes the $\sqrt{T}$-convergence rate and asymptotic normality of the exact SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ in (3.2) optimizing over *purely dimensional* sieves $\{\Theta_T\}_{T\in\mathbb{N}}$ (see Definition A.82 and Remarks A.83 and A.84). The simplification introduced by assuming that the sieves are *purely dimensional* w.r.t. $\{Q_{T,S}\}_{T\in\mathbb{N}}$ is well appreciated by noting that the term $\|\boldsymbol{\theta}_{T,S}^{*}-\boldsymbol{\theta}_{T,S}^{**}\|_{\Theta}$ in (3.7) vanishes.

Theorem 3.5.1. *Let Assumptions 3.1.1-3.5.8 hold. Furthermore, let* $\{\Theta_T\}_{T\in\mathbb{N}}$ *be purely dimensional sieves w.r.t. the sequence* $\{Q_{T,S}\}_{T\in\mathbb{N}}$*. Then, for fixed* $S\in\mathbb{N}$*, the exact SNPII estimator* $\hat{\boldsymbol{\theta}}_{T,S}$ *defined in (3.2) satisfies* $\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0\|=O_p(1)$ *as* $T\to\infty$ *and also,*

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0)\xrightarrow{d}-\text{inv}\Big(\nabla Q_{\infty}^{\nabla_{\mathbb{S}_{\Theta}}}(\boldsymbol{\theta}_0,\cdot)\Big)\big(\mathbb{G}_S(\boldsymbol{\theta}_0)\big)\ \ \text{as}\ \ T\to\infty. \tag{3.10}$$

Inspection of the proof of Theorem 3.5.1 reveals that an identical result can be immediately obtained for an approximate SNPII estimator $\hat{\boldsymbol{\theta}}_T$ satisfying the condition $\nabla_{\Theta_T}Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_{\Theta_T})=o_p(T^{-1/2})$. This approximate Z-estimator formulation can be easily obtained from the approximate extremum estimator defined in (3.3) under appropriate smoothness conditions.

Corollary 3.5.1. *Let Assumptions 3.1.1-3.5.8 hold. Furthermore, let* $\eta_T=o_p(T^{-1/2})$ *in (3.3) and* $\{\Theta_T\}_{T\in\mathbb{N}}$ *be purely dimensional sieves w.r.t. the sequence* $\{Q_{T,S}\}_{T\in\mathbb{N}}$*. Then, the approximate SNPII estimator* $\hat{\boldsymbol{\theta}}_{T,S}$ *in (3.3) satisfies* $\nabla_{\Theta_T}Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_{\Theta_T})=o_p(T^{-1/2})$ *and* $\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0\|=O_p(1)$ *and (3.10) as* $T\to\infty$*, for fixed* $S\in\mathbb{N}$.

When the sieves $\{\Theta_T\}_{T\in\mathbb{N}}$ are not assumed to be *purely dimensional* w.r.t. the sequence $\{Q_{T,S}\}_{T\in\mathbb{N}}$ then the appropriate $o_p(T^{-1/2})$ convergence rate of the term $\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta$ in (3.7) must be derived by other means. Assumption 3.5.9 below provides us with a sufficient condition for $\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta = o_p(T^{-1/2})$.

Assumption 3.5.9. *The sequence of directional derivatives* $\left\{\nabla\tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta},\boldsymbol{\theta}')\right\}_{T\in\mathbb{N}}$ *satisfies* $\nabla\tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta},\boldsymbol{\theta}') \xrightarrow{p} \nabla\boldsymbol{\beta}^*_i(\boldsymbol{\theta},\boldsymbol{\theta}')$ *as* $T\to\infty$ *for every* $(i,\boldsymbol{\theta},\boldsymbol{\theta}')\in\mathbb{N}\times\Theta\times\text{lin}\Theta$.

The following theorem establishes a $\sqrt{T}$-convergence rate for the exact SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ under the added influence of this assumption.

Theorem 3.5.2. *Let Assumptions 3.1.1-3.5.9 hold. Then, the exact SNPII estimator* $\hat{\boldsymbol{\theta}}_{T,S}$ *defined in (3.2) satisfies* $\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\| = O_p(1)$ *as* $T\to\infty$ *and the weak convergence in (3.10) for fixed* $S\in\mathbb{N}$.

Corollary 3.5.2. *Let Assumptions 3.1.1-3.5.9 hold and* $\eta_T = o_p(T^{-1/2})$ *in (3.3). Then the approximate SNPII estimator* $\hat{\boldsymbol{\theta}}_{T,S}$ *in (3.3) satisfies* $\nabla_{\Theta_T}Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_{\Theta_T}) = o_p(T^{-1/2})$ *and* $\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\| = O_p(1)$ *and (3.10) as* $T\to\infty$, *for fixed* $S\in\mathbb{N}$.

3.6 Statistical Inference with an Approximation of the Asymptotic Distribution

In non-trivial applications, it will be difficult to conduct statistical inference by considering in isolation the Theorems 3.5.1 and 3.5.2 or Corollaries 3.5.1 and 3.5.2 derived above. Indeed, a number of complications are likely to afflict anyone attempting to make use of the asymptotic distribution in (3.10). By relying on *(i)* the limit criterion μ_∞; *(ii)* the directional derivatives of Q_∞ w.r.t. the entire set of Schauder basis vectors $\mathbb{S}_\Theta$; and *(iii)* the Gaussian process $\mathbb{G}_S(\boldsymbol{\theta}_0)$; the weak convergence results above make use of an infinite system of estimating equations and depend on the asymptotic distribution of infinitely many auxiliary estimators. Dealing with the infinite system and the infinite vector of auxiliary estimators is, at the very minimum, unpractical. We are thus naturally led to ask the following question: *can we conduct inference based only on the asymptotic distribution of a finite number of auxiliary estimators and estimating equations?* Fortunately, the answer is yes.

Theorem 3.6.1 establishes the validity of a double asymptotic approximation argument where (3.10) is approximated by a distribution making use a finite number of auxiliary estimators and estimating equations. First, it makes use only of the asymptotic distribution $\pi_k\big(\mathbb{G}_S(\boldsymbol{\theta}_0)\big) \sim N(0,\Sigma_k(\boldsymbol{\theta}_0))$ of a finite subset of the infinite vector of auxiliary estimators (instead of the entire Gaussian process $\mathbb{G}_S(\boldsymbol{\theta}_0)$), where π_k denotes a projection from $\mathcal{B}$ into the first k auxiliary factor spaces $\mathcal{B}_1\times...\times\mathcal{B}_k$.

Second, it uses the criterion divergence μ_k that "concentrates" only on the relevant finite vector of auxiliary estimators in $\mathcal{B}_1 \times ... \times \mathcal{B}_k$, instead of μ_∞ that takes the entire sequence in $\mathcal{B}$ into account. Third, it makes use only of a finite set of estimating equations by taking the directional derivatives in the direction of the basis vectors of the relevant finite dimensional sieve $\mathbb{S}_{\Theta_k}$.

Theorem 3.6.1. (Approximation of Asymptotic Distribution) *Let Assumptions 3.1.1-3.5.8 hold. Then, Ψ_k converges weakly to $-\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot)\Big)(\mathbb{G}_S(\boldsymbol{\theta}_0))$ in (3.10) as $k \to \infty$, where,*

$$\Psi_k := -\text{inv}\Big(\nabla \mu_k^\nabla \Big(\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0)\,,\, \nabla \boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0, \cdot)\Big)\Big)\Big(\pi_k\Big(\mathbb{G}_S(\boldsymbol{\theta}_0)\Big)\Big)$$

with $\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0) := \Big(\,\Delta_\infty(\boldsymbol{\theta}_0)\,,\, \nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_k})\,\Big)$, *and also*

$$\nabla \boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0, \cdot) := \Big(\,\nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_k})\,,\, \nabla\Delta_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \cdot)\,\Big).$$

In applications, an especially interesting choice of the 'approximation variable' k is one that lets $\pi_k\Big(\mathbb{G}_S(\boldsymbol{\theta}_0)\Big)$ coincide with the asymptotic distribution of those auxiliary statistics that have been used in the estimation procedure. In the context of our reference Example 3.2.1, this consists of setting $k = k_T$. Indeed, the usefulness of Theorem 3.6.1 is well appreciated by turning back to our reference example.

Example 3.6.1. *Consider again the case of the wighted quadratic criterion divergence. Set k in Theorem 3.6.1 above equal to k_T. Then, Ψ_{k_T} can be shown to correspond to the usual asymptotic distribution obtained for the classical parametric indirect inference estimator in Gourieroux et al. (1993).*[24] *Obviously, one should not conclude from this that there exists no price to be paid for generality. Indeed, under the strict satisfaction of the correct specification axiom in Gourieroux et al. (1993), Ψ_{k_T} corresponds to the exact asymptotic distribution of $\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0)$. Here, it is only an approximation.*[25]

Finally, a word conducting inference on smooth functionals of $\hat{\boldsymbol{\theta}}_{T,S}$. Suppose that our interest lies in conducting inference on $\phi(\boldsymbol{\theta}_0)$ for some smooth functional $\phi : \Theta \to \Phi$. Then desired results follow naturally from application of an appropriate delta method.

[24] Estimation of several quantities such as the asymptotic variance-covariance matrix above is discussed also in Gourieroux and Monfort (1996).

[25] It is important to avoid any confusion concerning the fact that inference is focused on $\boldsymbol{\theta}_0$. Let $\boldsymbol{\theta}_0$ be a real-valued sequence. On finite samples, let the sieves restrict our attention to those sequences that are identically zero after a certain index value. Does inference on $\boldsymbol{\theta}_0$ require us to specify an infinite sequence? The answer is no. If $\boldsymbol{\theta}_0$ is believed to have non-zero entries after that index (i.e. if $\boldsymbol{\theta}_0$ is believed to lye outside the sieve) then, we shall reject any null that contradicts $\boldsymbol{\theta}_0 \in \Theta_k$. As such, null hypothesis of interest will consist only those under which $\boldsymbol{\theta}_0$ is also a sequence satisfying the uniform zero constraint (i.e. finite).

Corollary 3.6.1. *Let (3.10) hold true. Let $\phi : \Theta \to \Phi$ denote a Hadamard differentiable functional. Define $\mathbb{G}_S^* := -\text{inv}\Big(\nabla Q_\infty^{\nabla_S \Theta}(\boldsymbol{\theta}_0, \cdot)\Big)(\mathbb{G}_S(\boldsymbol{\theta}_0))$. Then, it follows by Lemma C.2.1 that,*

$$\sqrt{T}\Big(\phi(\hat{\boldsymbol{\theta}}_{T,S}) - \phi(\boldsymbol{\theta}_0)\Big) \xrightarrow{d} \nabla\phi\Big(\boldsymbol{\theta}_0, \mathbb{G}_S^*\Big) \;\; as \;\; T \to \infty.$$

3.7 Heterogeneity and Dependence

We have until now proceeded without mentioning explicitly the dependence and heterogeneity properties of both observed and simulated data. By avoiding such considerations we have been able to focus on what is essential to the argument: *the properties of the auxiliary estimators, the criterion divergence and the parameter space*. Indeed, there is strictly speaking, no need to elaborate further on the properties of the data as long as the conditions postulated above are satisfied. It is however clear that the assumed behavior of the auxiliary estimators $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ contains in itself statements about the DGP and the *postulated model*. In this sense, it is important to offer some remarks on how the desirable properties of auxiliary estimators might be obtained.

Since auxiliary estimators are allowed to take values on well-chosen, compact, finite dimensional sets. Their properties should in principle be easy to ascertain by appealing to the existing vast body literature on parametric M-estimators, Z-estimators, and others. Typical conditions for deriving the consistency and asymptotic normality of such estimators involve concepts of *L_p-approximation*, *near epoch dependence*, *weak or strong mixing*, and others. These conditions allow for various forms of heterogeneity and dependence. Relevant results can be found in Gallant and White (1988b), White (1994) and Pötscher and Prucha (1997) among others.

Having established the behavior of auxiliary estimators, one matter that has still been left unanswered concerns the convergence of the SNPII criterion function to a well defined limit. For concreteness let us rely one last time on our example of reference. At first sight, one might be troubled by the application of a Law of Large numbers for weighted averages over the infinite vectors of auxiliary statistics $\Delta_{T,S}(\boldsymbol{\theta}) := (\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}))$ whose heterogeneity and dependence is hard to understand. Fortunately however, the rate of convergence of $\Delta_{T,S}(\boldsymbol{\theta})$ to $\Delta_\infty(\boldsymbol{\theta})$ and the design of the weights $w_{T,i}$ can be called upon to play a decisive role. An unnecessarily strong albeit pleasantly simple set of sufficient conditions is the following (see Proposition A.89): *(i)* $\sup_{i\in\mathbb{N}} \|\boldsymbol{\beta}_i^*(\boldsymbol{\theta})\|_{\mathcal{B}_i} < \infty \, \forall \boldsymbol{\theta} \in \Theta$; and *(ii)* $w_{T,i} = \mathbb{1}(L < k_T)/2^i \, \forall (i,T) \in \mathbb{N} \times \mathbb{N}$. Various other conditions can be used. See e.g. Han and Phillips (2006) for scaling factors that can be used in ensuring that a weighted quadratic criterion function converges to an appropriate limit.

3.8 Uniform Convergence of Auxiliary Estimators

The consistency of the SNPII estimator derived in this chapter relied on the uniform convergence of auxiliary estimators over the index $i \in \mathbb{N}$ and across sieves $\{\Theta_T\}_{T\in\mathbb{N}} \subseteq \Theta$ (Assumption 3.4.2). This section offers some remarks on the verification of this assumption.

It is convenient to note first that uniformity in $\boldsymbol{\theta} \in \Theta$ is a common condition in indirect inference estimators; see e.g. Dridi and Renault (2000) and Nickl and Pötscher (2009). Furthermore, uniform convergence across sieves can be obtained by *(i)* imposing sufficient smoothness conditions on each auxiliary estimator $\tilde{\boldsymbol{\beta}}^i_{T,S}$, *(ii)* deriving the smoothness of the infinite vector $\tilde{\boldsymbol{\beta}}_{T,S}$ under the product topology and the already discussed results in Appendix C, and *(iii)* applying adapted versions (as noted by Chen (2007)) of the *generic uniform convergence* theorems of Newey (1991), Andrews (1987), Andrews (1992) and Potscher and Prucha (1989, 1994) (see Lemma 1.1.2). Uniformity across sieves is thus nothing new and should not cause any extra difficulties.

On the contrary, the uniform convergence of $\hat{\boldsymbol{\beta}}^i_{T,S}$ and $\tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta})$ over $i \in \mathbb{N}$ is uncommon to indirect inference estimators. This is so, simply because classical parametric indirect inference estimators make use of a single (or at least finitely many) auxiliary estimators. Hence, pointwise convergence typically implies uniform convergence. Now, there are two points worth mentioning in what concerns the uniform convergence of auxiliary estimators over $i \in \mathbb{N}$. First, as mentioned before, it is not a necessary condition (albeit a helpful one) for the results derived in chapter. Clearly, all that is needed is the uniform convergence across sieves. Uniformity over i could be relaxed to simple pointwise convergence (under the product topology) at the cost however of a considerably more complicated argument.

Second, uniform convergence over $i \in \mathbb{N}$ can nonetheless be obtained. The theory of *Empirical Processes* offers an especially wide range of conditions under which uniform convergence occurs over families of estimators. See e.g. the classical theory of empirical processes in Andrews (1986), Pollard (1989, 1990) and van der Vaart (1995) that is useful for auxiliary estimators taking the form of sample averages of nonlinear transformations of the data (e.g. sample moments), or Chebana (2007, 2009) for uniform convergence results specialized to hold over families (or infinite vectors) of M-estimators.

Finally, note that the theory of *generic uniform convergence* can also be easily applied to obtain uniform convergence jointly across sieves $\{\Theta_T\}$ and over $i \in \mathbb{N}$. Recall that the fundamental objective is that of obtaining the uniform convergence of $Q_{T,S}$ across the sieves which is implied by the uniform convergence of μ_T on $\mathcal{B}$ and the uniform convergence of $\tilde{\boldsymbol{\beta}}_{T,S}$ across sieves. Assumptions 3.8.1 and 3.8.2 make use

of pointwise convergence plus a generalized Hölder condition that implies *uniform asymptotic stochastic equicontinuity*, to obtain uniform convergence of μ_T and $\tilde{\boldsymbol{\beta}}_{T,S}$ across sieves in Θ and over $i \in \mathbb{N}$. Simple adaptation of appropriate theorems in Andrews (1992) and Davidson (1994, chp.21) yield the desired result.

Assumption 3.8.1. *The auxiliary estimators $\hat{\boldsymbol{\beta}}_T^i$ converge in probability to $\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)$ as $T \to \infty$ for every $i \in \mathbb{N}$ and the $\tilde{\boldsymbol{\beta}}_{T,s}^i$ converges in probability and pointwise on Θ to $\boldsymbol{\beta}_i^*(\boldsymbol{\theta})$ as $T \to \infty$ for every $i \in \mathbb{N}$. Furthermore, $\tilde{\boldsymbol{\beta}}_{T,s}^i$ satisfies the generalized Hölder condition*

$$\delta_{\mathcal{B}_i}\Big(\tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta}) - \tilde{\boldsymbol{\beta}}_{T,s}^i(\boldsymbol{\theta}')\Big) \leq \zeta_T \xi\Big(\delta_\Theta(\boldsymbol{\theta}, \boldsymbol{\theta}')\Big)$$

a.s. $\forall (\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta_T \times \Theta_T$ and every $T > T^*$, where ξ is a nonstochastic function satisfying $\lim_{x \to 0} \xi(x) = 0$ and ζ_T is a stochastic sequence satisfying either (i) $\zeta_T = O_p(1)$ *or* (ii) $\limsup_{T \in \mathbb{N}} \zeta_T < \infty$ *a.s..*

Assumption 3.8.2. *The sequence $\{\mu_T\}_{T \in \mathbb{N}}$ converges pointwise to μ_∞ on $\mathcal{B}$. Furthermore, μ_T satisfies the following generalized Hölder condition*

$$\Big|\mu_T^*(\boldsymbol{\beta}) - \mu_T^*(\boldsymbol{\beta}')\Big| \leq \xi_\mu\Big(\delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}')\Big)$$

for every $(\boldsymbol{\beta}, \boldsymbol{\beta}') \in \mathcal{B} \times \mathcal{B}$ and every $T \in \mathbb{N}$ where $\xi : \mathbb{R} \to \mathbb{R}$ is ζ_μ-homogeneous (see Definition A.53) and satisfies $\lim_{x \to 0} \xi(x) = 0$.

3.9 Optimal Convergence Rates

The results obtained in this chapter might seem very surprising at first. Indeed, looking at Assumptions 3.1.1, 3.5.1 and 3.5.3, one might be tempted to conclude that the $\sqrt{T}$-consistency (in norm) of the SNPII estimator applies to separable Banach spaces with a Schauder basis in general. This however, is not true. Indeed, some 'hidden' assumptions have contributed to obtaining the $\sqrt{T}$-consistency w.r.t. $\|\cdot\|_\Theta$ of the SNPII estimator.

Important results that provided some guidance on the theory of convergence rate of sieve estimators were initially made available by Charles Stone, in the very same year that Grenander introduced the method of sieves. In particular, Stone (1982) established from the outset, the important result that non-parametric estimators could do no better than achieve a $T^{-(p-m)/(2p+d)}$-convergence rate in L^q-norm ($0 < q < \infty$), and a $(T^{-1} \log T)^{-(p-m)/(2p+d)}$ in the sup-norm, when called to estimate the m-th derivative of a non-parametric p-differentiable regression function θ_0 of a d-dimensional variable x. Various other similar results followed. For example Yatracos (1985) established the relation between the convergence rate of minimum distance estimators and the entropy of the parameter space of interest.

Remark 3.9.1. *Note that typical $\sqrt{T}$-consistency results for nonparametric estimators refer to pointwise convergence and are often specific to restrictive infinite dimensional spaces.*

Often, sieve estimators achieve the optimal rates of Stone (1982). However, these results provide us also with a strict upper bound on the convergence rate that we can expect from an SNPI estimator on large infinite dimensional spaces. As such, the result of $\sqrt{T}$-convergence rate for the SNPII estimator, $\sqrt{T}\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|$ is certainly surprising (if not even paradoxical).

Remark 3.9.2. *The key to understanding the $\sqrt{T}$-convergence result is to note that a number of conditions have been imposed which 'indirectly' restrict the 'size' of the parameter space. These restrictions explain why the SNPII estimator seems to defy existing optimality theorems.*

The assumptions that imposed 'indirect' restrictions on Θ were those concerned with the properties of the binding function. Indeed, inspection of the conditions imposed in this chapter reveals that consistency alone requires already the parameter space to be homeomorphic to $\mathcal{B} \equiv \mathbb{R}^\infty$. Together with the maintained smoothness conditions used in obtaining $\sqrt{T}$-convergence, Θ is essentially required to be diffeomorphic (w.r.t. Hadamard differentiability) to $\mathbb{R}^\infty$. Hence, the $\sqrt{T}$-convergence results, far from requiring only that Θ be a separable Banach spaces with a Schauder basis, actually impose that Θ be 'nearly' diffeomorphic to $\mathbb{R}^\infty$.

Examples of parameter spaces satisfying such properties are easy to devise. For example, under appropriate regularity conditions, the space of analytic functions satisfies quite trivially such an assumption. This space still contains the important spaces of all polynomial, exponential and trigonometric functions. However, this space is certainly much smaller than many other separable Banach spaces. This should be present in applications.

Nonetheless, by noting that the convergence rate of the SNPII estimator is derived from that of the auxiliary estimators, one can nonetheless conclude that the SNPII estimator achieves the optimal convergence rates of Stone (1982). Indeed, by making use of the appropriate sieve estimator as the auxiliary statistics of the SNPII estimator, the same level of generality and same convergence rates can always be obtained.

3.10 Conclusion

This chapter derived the consistency, $\sqrt{T}$-convergence and asymptotic Gaussianity of an SNPII estimator relying on an infinite number of auxiliary parametric estimators. Compared to Chapter 2, the results were derived in considerably more

detail. In particular, the assumptions used in this chapter have been traced back to primitive conditions on the criterion divergence and individual auxiliary estimators.

These results in this chapter not only confer important generality to indirect inference estimators, they obtain $\sqrt{T}$-convergence and asymptotic Gaussianity under settings allowing for greater data dependence and heterogeneity than generally found in the literature of sieve estimation. Under simple regularity conditions, the consistency, $\sqrt{T}$-convergence and asymptotic Gaussianity of functionals of the SNPII estimator was also derived as corollaries of the main theorems.

In the next chapter we finally, take a look at the finite-sample properties of the SNPII estimator. As we shall see, the Monte Carlo study that follows seems to confirm the theoretical properties derived in this chapter for the SNPII estimator.

3.11 Proofs

Proof of Theorem 3.3.1

Proof. Clearly, the $\mathcal{F}/\mathfrak{B}(\mathcal{B}_i)$-measurability of each auxiliary estimator $\tilde{\boldsymbol{\beta}}^i_{T,s}(\cdot,\boldsymbol{\theta}) : \Omega \to \mathcal{B}_i$ for every $(\boldsymbol{\theta},T,s,i) \in \Theta \times \mathbb{N} \times \{1,...,S\} \times \mathbb{N}$, postulated in Assumption 3.3.1, implies the $\mathcal{F}/\mathfrak{B}(\mathcal{B}_i)$-measurability of the average $\tilde{\boldsymbol{\beta}}^i_{T,S}(\cdot,\boldsymbol{\theta}) : \Omega \to \mathcal{B}_i$ obtained as $\tilde{\boldsymbol{\beta}}^i_{T,S}(\cdot,\boldsymbol{\theta}) = 1/S \sum_{s=1}^{S} \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta})$, for every $(\boldsymbol{\theta},T,S,i) \in \Theta \times \mathbb{N} \times \mathbb{N} \times \mathbb{N}$ by the continuity of vector addition and scalar multiplication under Assumption 3.1.3 (see Definition A.28) and measurability of continuous functions (Lemma A.11 and Corollary A.13). This implies in turn that, given Assumption 3.1.3, $\tilde{\boldsymbol{\beta}}_{T,S}(\cdot,\boldsymbol{\theta}) : \Omega \to \mathcal{B}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{B})$-measurable $\forall(\boldsymbol{\theta},T,S) \in \Theta \times \mathbb{N} \times \mathbb{N}$ (Lemma A.15 and Corollary A.27). By the same argument, the $\mathcal{F}/\mathfrak{B}(\mathcal{B}_i)$-measurability of the auxiliary estimators $\hat{\boldsymbol{\beta}}^i_T : \Omega \to \mathcal{B}_i \, \forall (T,i) \in \mathbb{N} \times \mathbb{N}$ implies the $\mathcal{F}/\mathfrak{B}(\mathcal{B})$-measurability of $\hat{\boldsymbol{\beta}}_T : \Omega \to \mathcal{B} \, \forall T \in \mathbb{N}$. The continuity of vector addition and scalar multiplication yields the $\mathcal{F}/\mathfrak{B}(\mathcal{B})$-measurability of $\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\cdot,\boldsymbol{\theta})$. Furthermore, Assumption 3.3.3 implies that $\mu_T : \mathcal{B} \to \mathbb{R}$ is $\mathfrak{B}(\mathcal{B})/\mathfrak{B}(\mathbb{R})$-measurable (Corollary A.13) for every $T \in \mathbb{N}$ (Lemma A.11 and Corollary A.13), and hence, together with the measurability of $\tilde{\boldsymbol{\beta}}_{T,S}(\cdot,\boldsymbol{\theta}) : \Omega \to \mathcal{B} \, \forall(\boldsymbol{\theta},T,S) \in \Theta \times \mathbb{N} \times \mathbb{N}$ and $\hat{\boldsymbol{\beta}}_T \, \forall T \in \mathbb{N}$, we have that $Q_{T,S}(\boldsymbol{\theta}) := \mu_T\Big(\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) : \Omega \to \mathbb{R}$ is $\mathcal{F}/\mathfrak{B}(\mathbb{R})$-measurable for every $(\boldsymbol{\theta},T,S) \in \Theta \times \mathbb{N} \times \mathbb{N}$ by measurability of measurable compositions (Lemma A.14).

Now, Assumption 3.3.2 implies immediately the continuity of the average map $\tilde{\boldsymbol{\beta}}^i_{T,S}(\omega,\cdot) : \Theta \to \mathcal{B}_i$ on $\Theta \;\; \forall(\omega,T,S,i) \in \Omega \times \mathbb{N} \times \mathbb{N} \times \mathbb{N}$ (under Assumption 3.1.3, Definition A.28 and Lemma A.29). This in turn implies (under Assumption 3.1.3) the continuity of $\tilde{\boldsymbol{\beta}}_{T,S}(\omega,\cdot) : \Theta \to \mathcal{B}$ on $\Theta \;\; \forall(\omega,T,S) \in \Omega \times \mathbb{N} \times \mathbb{N}$ (Lemma A.18). Together with the continuity of μ_T postulated in Assumption 3.3.3 this implies the continuity of $Q_{T,S}(\omega,\cdot) := \mu_T\Big(\hat{\boldsymbol{\beta}}_T(\omega) - \tilde{\boldsymbol{\beta}}_{T,S}(\omega,\cdot)\Big) : \Theta \to \mathbb{R}$ on Θ for every

$(\omega, T, S) \in \Omega \times \mathbb{N} \times \mathbb{N}$ (Lemma A.29).

Finally, $\mathcal{F}/\mathfrak{B}(\mathbb{R})$-measurability of $Q_{T,S}(\boldsymbol{\theta}) : \Omega \to \mathbb{R}$ for every $(\boldsymbol{\theta}, T, S) \in \Theta \times \mathbb{N} \times \mathbb{N}$ and continuity of $Q_{T,S}(\omega, \cdot) : \Theta \to \mathbb{R}$ on Θ for every $(\omega, T, S) \in \Omega \times \mathbb{N} \times \mathbb{N}$ implies by Lemma A.30 that $Q_{T,S} : \Omega \times \Theta \to \mathbb{R}$ is $\mathcal{F} \otimes \mathfrak{B}(\Theta)/\mathfrak{B}(\mathbb{R})$-measurable. Together with Assumptions 3.1.1 and 3.1.2 the desired result follows from Lemmas A.31, A.32 and Corollary A.33 adapted from Debreu (1967, Theorem 4.5), Hildenbrand (1974, p.55) and White and Wooldrige (1991, Theorem 2.2, p.646), i.e. that there exists a $\hat{\boldsymbol{\theta}}_{T,S} : \Omega \to \Theta_T$ satisfying (3.3) for every $T \in \mathbb{N}$ and $S \in \mathbb{N}$ that is $\mathcal{F}/\mathfrak{B}(\Theta_T)$-measurable.[26] □

Proof of Theorem 3.4.1

Proof. Note first that, given Assumptions 3.1.1, 3.1.3 and 3.4.3 the product binding function $\boldsymbol{\beta}^* : \Theta \to \mathcal{B}$ is a homeomorphism and thus injective, continuous and open (Proposition A.48). By injectivity $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)$ and $\boldsymbol{\beta}_0^* \neq \boldsymbol{\beta}^*(\boldsymbol{\theta}) \; \forall \boldsymbol{\theta} \in \Theta \backslash \{\boldsymbol{\theta}_0\}$. By the properties of divergences (Definition A.51) that,

$$Q_\infty(\boldsymbol{\theta}_0) := \mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\Big) = \mu_\infty\big(0\big) = 0,$$

and

$$Q_\infty(\boldsymbol{\theta}) := \mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) > 0 \; \forall \boldsymbol{\theta} \in \Theta \backslash \{\boldsymbol{\theta}_0\}.$$

By openness of $\boldsymbol{\beta}^*$ we have that $\boldsymbol{\beta}^*\Big(S(\boldsymbol{\theta}_0, \epsilon)\Big)$ is an open subset of $\mathcal{B}$ containing $\boldsymbol{\beta}_0^* := \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)$ for every open ball $S(\boldsymbol{\theta}_0, \epsilon) \subset \Theta$, $\epsilon > 0$ centered at $\boldsymbol{\theta}_0$. Hence, for any $\epsilon > 0$, the set

$$S_{\mathcal{B}} := \Big\{\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta}) \; , \; \boldsymbol{\theta} \in S(\boldsymbol{\theta}_0, \epsilon)\Big\}$$

is an open subset of $\mathcal{B}$ containing the origin $0_{\mathcal{B}}$ of $\mathcal{B}$. As a result, there exist an open ball $S(0_{\mathcal{B}}, \epsilon')$ of radius $\epsilon' > 0$, centered at the origin $0_{\mathcal{B}} \in \mathcal{B}$, such that $S(0_{\mathcal{B}}, \epsilon') \subset S_{\mathcal{B}} \subset \mathcal{B}$. Furthermore, their complements in $\mathcal{B}$ satisfy $S_{\mathcal{B}}^c \subset S^c(0_{\mathcal{B}}, \epsilon') \subset \mathcal{B}$. Together with Assumption 3.4.5 it thus follows immediately that $\boldsymbol{\theta}_0$ is identifiably unique since for every $\epsilon > 0$, there exist $\epsilon' > 0$ such that,

$$\begin{aligned}
\inf_{\boldsymbol{\theta} \in S^c_{\boldsymbol{\theta}_0}(\epsilon)} \Big|Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)\Big| &= \inf_{\boldsymbol{\theta} \in S^c(\boldsymbol{\theta}_0, \epsilon)} \Big|\mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\Big)\Big| \\
&= \inf_{\boldsymbol{\theta} \in S^c(\boldsymbol{\theta}_0, \epsilon)} \Big|\mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big)\Big| = \inf_{\boldsymbol{\beta} \in \boldsymbol{\beta}^*(S^c(\boldsymbol{\theta}_0, \epsilon))} \Big|\mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}\Big)\Big| \quad (3.11) \\
&= \inf_{\boldsymbol{\beta} \in S^c_{\mathcal{B}}} \Big|\mu_\infty(\boldsymbol{\beta})\Big| \geq \inf_{\boldsymbol{\beta} \in S^c(0_{\mathcal{B}}, \epsilon')} \Big|\mu_\infty\Big(\boldsymbol{\beta}_0^* - \boldsymbol{\beta}\Big)\Big| > 0,
\end{aligned}$$

[26] These results rely on the fact that under Assumption 3.1.1, Θ is a *Polish* space (Definition A.2). Note how Lemmas A.31 and A.32 allow for random sieves to be considered. Note also that, in what near-measurability is concerned, completeness and separability of Θ could be weakened to the requirement that Θ be a Souslin measurable space; see Stinchcombe and White (1992).

where the second equality follows by identity of indiscernibles of divergences (Definition A.51) and the last inequality by Assumption 3.4.5.

Now, given the Lipschitz weakness of $\delta_{\mathcal{B}}$ postulated in Assumption 3.4.1, the uniform convergence of $\tilde{\boldsymbol{\beta}}^i_{T,s}$ in i and $\boldsymbol{\theta}$ (Assumption 3.4.2) implies the uniform convergence of the product empirical binding function $\boldsymbol{\beta}^*$ on Θ. Indeed, for every $\epsilon > 0$, it holds true that,

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\,,\, \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) > \epsilon \Big) &\leq \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} k \cdot \sup_{i\in\mathbb{N}} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big) \\
&= \mathbb{P}\Big(k \cdot \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big) \\
&= \mathbb{P}\Big(k \cdot \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \Big\| 1/S \sum_{s=1}^{S} \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta}) - 1/S \sum_{s=1}^{S} \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big) \\
&\leq \mathbb{P}\Big(k \cdot \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} 1/S \sum_{s=1}^{S} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big) \\
&\leq \mathbb{P}\Big(k/S \sum_{s=1}^{S} \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big)
\end{aligned} \tag{3.12}$$

for every $(T, S) \in \mathbb{N} \times \mathbb{N}$ and some $k \in \mathbb{R}^+$, and where the first inequality follows by Assumption 3.4.1, the second by norm sub-additivity, and the third by supremum sub-additivity. Hence, by Assumption 3.1.3 the continuous mapping Theorem (Corollary A.55, see also Definition A.28 and note that a degenerate random variable is separable) and part *(ii)* of Assumption 3.4.2 we have that, for every $\epsilon > 0$,

$$\lim_{T\to\infty} \mathbb{P}\Big(k/S \sum_{s=1}^{S} \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big) = 0.$$

This implies by (3.12) and Lemma A.56 that,

$$\lim_{T\to\infty} \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\,,\, \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) > \epsilon \Big) = 0 \ \forall \epsilon > 0. \tag{3.13}$$

For almost sure uniform convergence simply note that following the argument in (3.12) and Lemma A.56,

$$\lim_{T\to\infty} \sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\,,\, \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) \ \leq \ \lim_{T\to\infty} k/S \sum_{s=1}^{S} \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i},$$

and hence, since by Assumption 3.1.3 the continuous mapping Theorem (Corollary A.55) and part *(ii)* of Assumption 3.4.2,

$$\mathbb{P}\Big(\lim_{T\to\infty} k/S \sum_{s=1}^{S} \sup_{\boldsymbol{\theta}\in\Theta_T} \sup_{i\in\mathbb{N}} \Big\| \tilde{\boldsymbol{\beta}}^i_{T,s}(\boldsymbol{\theta}) - \boldsymbol{\beta}^*_i(\boldsymbol{\theta}) \Big\|_{\mathcal{B}_i} > \epsilon \Big) = 0 \ \forall \epsilon > 0,$$

we have by Lemma A.56,[27]

$$\mathbb{P}\Big(\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\,,\,\boldsymbol{\beta}^*(\boldsymbol{\theta})\Big)>\epsilon\Big)=0\,\forall\,\epsilon>0. \tag{3.14}$$

Convergence in probability and a.s. of $\hat{\boldsymbol{\beta}}_T$ is implied by Assumption 3.4.1, part *(i)* of Assumption 3.4.2 and Lemma A.56 since it follows immediately from $\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T\,,\,\boldsymbol{\beta}_0^*\big)\leq k\cdot\sup_{i\in\mathbb{N}}\big\|\hat{\boldsymbol{\beta}}_T^i-\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)\big\|_{\mathcal{B}_i}$ that,

$$\lim_{T\to\infty}\mathbb{P}\Big(\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T\,,\,\boldsymbol{\beta}_0^*\big)>\epsilon\Big)\leq\lim_{T\to\infty}\mathbb{P}\Big(k\cdot\sup_{i\in\mathbb{N}}\big\|\hat{\boldsymbol{\beta}}_T^i-\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)\big\|_{\mathcal{B}_i}>\epsilon\Big)=0\;\;\forall\,\epsilon>0, \tag{3.15}$$

and also that,

$$\mathbb{P}\Big(\lim_{T\to\infty}\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T\,,\,\boldsymbol{\beta}_0^*\big)>\epsilon\Big)\leq\mathbb{P}\Big(\lim_{T\to\infty}k\cdot\sup_{i\in\mathbb{N}}\big\|\hat{\boldsymbol{\beta}}_T^i-\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0)\big\|_{\mathcal{B}_i}>\epsilon\Big)=0\;\;\forall\,\epsilon>0. \tag{3.16}$$

Uniform convergence in probability of the centered empirical binding function $\Delta_{T,S}(\boldsymbol{\theta}):=\hat{\boldsymbol{\beta}}_T-\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ to $\Delta_\infty(\boldsymbol{\theta}):=\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)-\boldsymbol{\beta}^*(\boldsymbol{\theta})$ across the sequence of sieves $\{\Theta_T\}_{T\in\mathbb{N}}$ now follows immediately from (3.13) and (3.15) since it holds true that,

$$\begin{aligned}
\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\Delta_{T,S}(\boldsymbol{\theta})\,,\,\Delta_\infty(\boldsymbol{\theta})\big)>\epsilon\Big)&=\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\Delta_{T,S}(\boldsymbol{\theta})\,,\,\Delta_\infty(\boldsymbol{\theta})\big)>\epsilon\Big)\\
&=\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T-\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\,,\,\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)-\boldsymbol{\beta}^*(\boldsymbol{\theta})\big)>\epsilon\Big)\\
&=\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T-\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})+\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})-\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\,,\\
&\qquad\qquad\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)-\boldsymbol{\beta}^*(\boldsymbol{\theta})+\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})-\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\big)>\epsilon\Big)\\
&\leq\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\Big[\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\big)+\delta_{\mathcal{B}}\big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})-\boldsymbol{\beta}^*(\boldsymbol{\theta})\big)\Big]>\epsilon\Big)\\
&\leq\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\big)+\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})-\boldsymbol{\beta}^*(\boldsymbol{\theta})\big)>\epsilon\Big)\\
&=\mathbb{P}\Big(\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\big)+\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})-\boldsymbol{\beta}^*(\boldsymbol{\theta})\big)>\epsilon\Big)\\
&\leq\mathbb{P}\Big(\delta_{\mathcal{B}}\big(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\big)>\epsilon/2\Big)\\
&\quad+\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})-\boldsymbol{\beta}^*(\boldsymbol{\theta})\big)>\epsilon/2\Big)
\end{aligned} \tag{3.17}$$

[27] It is also clear that under appropriate regularity conditions, the almost sure convergence of the product binding function $\boldsymbol{\beta}^*$ uniformly on Θ is also obtained directly by convergence of the projection maps under Assumption 3.1.3 and Corollary A.16 without the need for Assumption 3.4.1.

where the second equality follows from the fact that the product metric $\delta_{\mathcal{B}}$ inherits the translation invariance from the norms $\|\cdot\|_{\mathcal{B}_i}$ for every i, the first inequality follows from metric sub-additivity, the second by sub-additivity of the supremum, and the third by the fact that $\{a+b>\epsilon\} \subseteq \{a>\epsilon/2\} \cup \{b>\epsilon/2\}$ and that, for random events, this implies $\mathbb{P}(a+b>\epsilon) \leq \mathbb{P}(a>\epsilon/2) + \mathbb{P}(b>\epsilon/2)$. Finally, by the convergence results obtained in (3.13) and (3.15), the last two terms converge to zero which implies by Lemma A.56 the convergence of the centered empirical binding function,

$$\lim_{T\to\infty} \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta}),\, \Delta_\infty(\boldsymbol{\theta})\Big) > \epsilon \Big) = 0 \;\; \forall \epsilon > 0. \tag{3.18}$$

The almost sure counterpart of this result is obtained by following the same argument as in (3.17) to conclude that, for every $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\Big(\lim_{T\to\infty} \sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta}), \Delta_\infty(\boldsymbol{\theta})\Big) > \epsilon \Big) &\leq \mathbb{P}\Big(\lim_{T\to\infty} \delta_{\mathcal{B}}\Big(\hat{\boldsymbol{\beta}}_T\,,\; \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\Big) > \epsilon/2 \Big) \\ &\quad + \mathbb{P}\Big(\lim_{T\to\infty} \sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\,,\; \boldsymbol{\beta}^*(\boldsymbol{\theta})\Big) > \epsilon/2 \Big)\end{aligned}$$

and thus obtain by the a.s. convergence results in (3.14) and (3.16),

$$\mathbb{P}\Big(\lim_{T\to\infty} \sup_{\boldsymbol{\theta}\in\Theta_T} \delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta}),\, \Delta_\infty(\boldsymbol{\theta})\Big) > \epsilon \Big) = 0 \;\; \forall \epsilon > 0. \tag{3.19}$$

Now, the uniform convergence across $\{\Theta_T\}_{T\in\mathbb{N}}$ of $Q_{T,S}(\boldsymbol{\theta}) := \mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)$ is obtained by noting that, for every $T \in \mathbb{N}$ and every $\epsilon > 0$, it holds true that,

$$\begin{aligned}\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| Q_{T,S}(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}) \Big| > \epsilon \Big) &= \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big) \Big| > \epsilon \Big) \\ &\leq \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) \Big| \\ &\qquad + \sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big) \Big| > \epsilon \Big) \\ &\leq \mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) \Big| > \epsilon/2 \Big) \\ &\qquad + P\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big) \Big| > \epsilon/2 \Big)\end{aligned} \tag{3.20}$$

where the first inequality is obtained by simply adding and subtracting $\mu_\infty(\Delta_{T,S}(\boldsymbol{\theta}))$ and by norm sub-additivity of the absolute value and supremum functions. Now,

$$\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) \Big| > \epsilon/2 \Big) \to 0 \tag{3.21}$$

holds true by the sure uniform convergence of μ_T (Assumption 3.4.4), and

$$\mathbb{P}(\sup_{\boldsymbol{\theta}\in\Theta_T} \Big| \mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big) \Big| > \epsilon/2 \Big) \to 0 \tag{3.22}$$

is implied by the uniform convergence of the centered empirical binding function in (3.18) and by uniform continuity of μ_∞ on compact sets (Assumption 3.4.4 and Heine-Cantor Theorem in Lemma A.58).

Let us start by noticing that the closed metric ϵ'-ball centered at the zero element $0_{\mathcal{B}}$ of $\mathcal{B}$ and defined as $\bar{S}(0_{\mathcal{B}}, \epsilon') := \left\{\boldsymbol{\beta} \in \mathcal{B} : \delta_{\mathcal{B}}(\boldsymbol{\beta}, 0_{\mathcal{B}}) \leq \epsilon'\right\}$ is compact in the product topology, for every $\epsilon' > 0$. This follows by observing that, for every $\epsilon' > 0$, there exists a compact set that contains $\bar{S}(0_{\mathcal{B}}, \epsilon')$. Indeed, define,

$$\mathcal{B}_0(\epsilon'') := \left\{\boldsymbol{\beta} \in \mathcal{B} \ : \ \|\pi_i(\boldsymbol{\beta})\|_{\mathcal{B}_i} \leq \epsilon'' \, \forall \, i \in \mathbb{N}\right\}.$$

It is easy to verify that for every ϵ' there exists an ϵ'' such that $\bar{S}(0_{\mathcal{B}}, \epsilon') \subseteq \mathcal{B}_0(\epsilon'')$. Furthermore, compactness of $\mathcal{B}_0(\epsilon'')$ for every $\epsilon'' > 0$ follows naturally by noting that,

$$\mathcal{B}_0(\epsilon'') = \times_{i \in \mathbb{N}} \mathcal{B}_0^i(\epsilon'') \quad \text{where} \quad \mathcal{B}_0^i(\epsilon'') := \left\{\boldsymbol{\beta}_i \in \mathcal{B}_i \ : \ \|\boldsymbol{\beta}_i\|_{\mathcal{B}_i} \leq \epsilon''\right\}.$$

Since $\mathcal{B}_0^i(\epsilon'')$ is compact for every $i \in \mathbb{N}$, compactness of the product $\mathcal{B}_0(\epsilon'')$ follows from Tychonoff's Theorem (Lemma A.19). Finally, since for every $\epsilon > 0$, $\bar{S}(0_{\mathcal{B}}, \epsilon)$ is a closed subset of a compact set $\mathcal{B}_0(\epsilon'')$, for some $\epsilon' > 0$, then $\bar{S}(0_{\mathcal{B}}, \epsilon)$ is also compact. Now since $\mu_\infty : \mathcal{B} \to \mathbb{R}$ is continuous on $\mathcal{B}$, and $\bar{S}(0_{\mathcal{B}}, \epsilon')$ is compact $\forall \, \epsilon' > 0$, then μ_∞ it is uniformly continuous on $\bar{S}(0_{\mathcal{B}}, \epsilon') \, \forall \epsilon' > 0$, by the Heine-Cantor Theorem. As a result, for every $\epsilon > 0$, there exists an $\epsilon' > 0$, such that every $(\boldsymbol{\beta}, \boldsymbol{\beta}') \in \bar{S}(0, \epsilon') \times \bar{S}(0, \epsilon')$ having $\delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}') \leq \epsilon''$ satisfies $|\mu_\infty(\boldsymbol{\beta}) - \mu_\infty(\boldsymbol{\beta}')| < \epsilon$.

Let Θ^* be a compact subset of Θ and define a pair of maps $\boldsymbol{\beta} : \Theta \to \mathcal{B}$ and $\boldsymbol{\beta}' : \Theta \to \mathcal{B}$ that are continuous on Θ. By continuity, both $\boldsymbol{\beta}(\Theta^*)$ and $\boldsymbol{\beta}'(\Theta^*)$ are compact. The set $\mathcal{B}_{\Theta^*} := \boldsymbol{\beta}(\Theta^*) \times \boldsymbol{\beta}'(\Theta^*)$ is also compact. Since μ_∞ is continuous on $\mathcal{B}$, it is uniformly continuous on $\mathcal{B}_{\Theta^*} \subset \mathcal{B}$. For $\epsilon > 0$, $\exists \epsilon'' > 0$ such that,

$$\sup_{\boldsymbol{\theta} \in \Theta^*} \delta_{\mathcal{B}}\Big(\boldsymbol{\beta}(\boldsymbol{\theta}) \ , \ \boldsymbol{\beta}'(\boldsymbol{\theta})\Big) < \epsilon' \quad \Rightarrow \quad \sup_{\boldsymbol{\theta} \in \Theta^*} \Big|\mu_\infty\Big(\boldsymbol{\beta}(\boldsymbol{\theta})\Big) - \mu_\infty\Big(\boldsymbol{\beta}'(\boldsymbol{\theta})\Big)\Big| < \epsilon.$$

For every $\omega \in \Omega$, let us now define the set,

$$B^{\Delta}_{\Theta_T}(\omega) := \left\{\boldsymbol{\beta} \in \mathcal{B} \ : \ \Delta_{T,S}(\omega, \boldsymbol{\theta}) - \Delta_\infty(\boldsymbol{\theta}) \ , \ \boldsymbol{\theta} \in \Theta_T\right\}.$$

By compactness of $\Theta_T \, \forall T \in \mathbb{N}$ and continuity of $\Delta_{T,S}$ and Δ_∞ (derived in Theorem 3.3.1), it follows that $B^{\Delta}_{\Theta_T}(\omega)$ is compact for every $(\omega, T, S) \in \Omega \times \mathbb{N} \times \mathbb{N}$. Since a compact set is totally bounded and totally bounded sets are bounded, it follows that for every for every $(\omega, T, S) \in \Omega \times \mathbb{N} \times \mathbb{N}$, there exists an $\epsilon' > 0$ such that $B_{\Delta_T}(\omega) \subseteq \bar{S}(0, \epsilon')$. In other words,

$$\sup_{\boldsymbol{\theta} \in \Theta_T} \delta_{\mathcal{B}}\Big(\Delta_{T,S}(\omega, \boldsymbol{\theta}) \ , \ \Delta_\infty(\boldsymbol{\theta})\Big) < \epsilon'.$$

By uniform continuity of μ_∞ on $\bar{S}(0,\epsilon')$, this implies naturally that,

$$\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_\infty\Big(\Delta_{T,S}(\omega,\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\omega,\boldsymbol{\theta})\Big)\Big|<\epsilon.$$

As a result it follows that for every $\epsilon>0$, $\exists\,\epsilon'>0$ such that,

$$\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big)\Big|<\epsilon\Big)\geq\mathbb{P}\left(B_{\Delta_T}\subseteq\bar{S}(0,\epsilon')\right).$$

Finally, note that, for every $\omega\in\Omega$ and every $T\in\mathbb{N}$, having

$$\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\Delta_{T,S}(\omega,\boldsymbol{\theta}),\Delta_\infty(\boldsymbol{\theta})\Big)<2\epsilon'$$

implies by construction that $B_{\Delta_T}(\omega)\subseteq\bar{S}(0,\epsilon')$. Hence, $\forall T\in\mathbb{N}$ we have that,

$$\mathbb{P}\Big(B_{\Delta_T}\subseteq\bar{S}(0,\epsilon')\Big)\geq\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta})\,,\,\Delta_\infty(\boldsymbol{\theta})\Big)<2\epsilon'\Big).$$

The two previous inequalities can now be used to conclude that, for every $T\in\mathbb{N}$ and every $\epsilon>0$ there exists $\epsilon'>0$ such that,

$$\begin{aligned}\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big)\Big|<\epsilon\Big)&\geq\mathbb{P}\Big(B_{\Delta_T}\subseteq\bar{S}(0,\epsilon')\Big)\\&\geq\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta}),\Delta_\infty(\boldsymbol{\theta})\Big)<2\epsilon'\Big).\end{aligned}$$

As a result, (3.18) and Assumption 3.4.4 implies (3.22). In particular, since

$$\lim_{T\to\infty}\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta}),\Delta_\infty(\boldsymbol{\theta})\Big)<2\epsilon'\Big)=1\quad\forall\,\epsilon'>0,$$

it holds that,

$$\lim_{T\to\infty}\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big)\Big|<\epsilon\Big)=1\quad\forall\epsilon>0.$$

Finally, (3.21) and (3.22) imply by (3.20) that

$$\lim_{T\to\infty}\mathbb{P}\Big(\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|Q_{T,S}(\boldsymbol{\theta})-Q_\infty(\boldsymbol{\theta})\Big|>\epsilon\Big)=0\quad\forall\,\epsilon>0.\tag{3.23}$$

The almost sure counterpart of this result is obtained by the same argument. In particular, similarly to (3.20),

$$\begin{aligned}\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|Q_{T,S}(\boldsymbol{\theta})-Q_\infty(\boldsymbol{\theta})\Big|\leq&\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)\Big|\\&+\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big)\Big|\end{aligned}\tag{3.24}$$

and then,

$$\mathbb{P}\Big(\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|\mu_T\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)\Big|>\epsilon/2\Big)=0\,\forall\,\epsilon>0\tag{3.25}$$

holds also true by the sure uniform convergence of μ_T (Assumption 3.4.4), and

$$\mathbb{P}(\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\left|\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big)\right|>\epsilon/2\Big)=0\;\forall\,\epsilon>0 \tag{3.26}$$

is implied by the uniform convergence in (3.19) and uniform continuity of μ_∞ on compact sets by using once more the fact that uniform continuity preserves uniform convergence (Proposition A.57). In particular, the result is obtained in a similar way by noting that, for every $(\omega,S)\in\Omega\times\mathbb{N}$ and every $\epsilon>0$, $\exists\epsilon'>0$ such that,

$$\mathbb{P}\Big(\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\left|\mu_\infty\Big(\Delta_{T,S}(\boldsymbol{\theta})\Big)-\mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta})\Big)\right|<\epsilon\Big)\geq\mathbb{P}\left(\lim_{T\to\infty}B_{\Delta_T}\subseteq\bar{S}(0,\epsilon')\right).$$

Now, for every $(\omega,T,S)\in\Omega\times\mathbb{N}\times\mathbb{N}$, having

$$\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\Delta_{T,S}(\omega,\boldsymbol{\theta}),\Delta_\infty(\boldsymbol{\theta})\Big)<2\epsilon'$$

implies by construction that $\lim_{T\to\infty}B_{\Delta_T}(\omega)\subseteq\bar{S}(0,\epsilon')$. Hence, $\forall T\in\mathbb{N}$ we have that,

$$\mathbb{P}\Big(\lim_{T\to\infty}B_{\Delta_T}\subseteq\bar{S}(0,\epsilon')\Big)\geq\mathbb{P}\left(\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\delta_{\mathcal{B}}\Big(\Delta_{T,S}(\boldsymbol{\theta}),\Delta_\infty(\boldsymbol{\theta})\Big)<2\epsilon'\right).$$

The two previous inequalities can now be combined with (3.19) and Assumption 3.4.4 to obtain (3.26). Finally, (3.25) and (3.26) imply by (3.24) that

$$\mathbb{P}\Big(\lim_{T\to\infty}\sup_{\boldsymbol{\theta}\in\Theta_T}\Big|Q_{T,S}(\boldsymbol{\theta})-Q_\infty(\boldsymbol{\theta})\Big|>\epsilon\Big)=0\;\;\forall\,\epsilon>0. \tag{3.27}$$

Continuity of the limit criterion function Q_∞ on Θ follows from *(i)* the continuity of the product binding function $\boldsymbol{\beta}^*$ on Θ (implied by Assumptions 3.1.1, 3.1.3, 3.4.3 and Lemma A.46 since a homeomorphism is continuous by definition), *(ii)* the continuity of μ_∞ on $\mathcal{B}$ (Assumption 3.4.4), and *(iii)* the continuity of continuous compositions (Lemma A.29).

$$Q_\infty(\cdot):=\mu_\infty\Big(\boldsymbol{\beta}_0^*-\boldsymbol{\beta}^*(\cdot)\Big):\Theta\to\mathbb{R}\text{ is continuous in }\boldsymbol{\theta}\in\Theta. \tag{3.28}$$

Finally, recall that the measurability of $\hat{\boldsymbol{\theta}}_{T,S}$ follows from 3.1.1-3.3.3 and Theorem 3.3.1. Given Assumptions 3.1.1-3.4.5 and the intermediate results of *(i)* identifiable uniqueness of $\boldsymbol{\theta}_0$ obtained in (3.11), *(ii)* uniform convergence in probability of the criterion function $Q_{T,S}$ established in (3.23) and *(iii)* the continuity of the limit criterion function Q_∞ derived in (3.28); the desired conclusion that the approximate SNPII estimator $\hat{\boldsymbol{\theta}}_{T,S}$ defined in (3.3) and (3.4) satisfies $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S},\boldsymbol{\theta}_0)\xrightarrow{p}0$ follows by Lemma A.59 adapted from Theorem 3.1 in Chen (2007) (see also Proposition 2.4 and Corollary 2.6 in White and Wooldrige (1991)). The convergence $\delta_\Theta(\hat{\boldsymbol{\theta}}_{T,S},\boldsymbol{\theta}_0)\xrightarrow{a.s.}0$ follows by the same conditions and the uniform a.s. convergence of the criterion function $Q_{T,S}$ established in (3.27) and Lemma A.59 (see Theorem 3.1 and Remark 3.2 in Chen (2007)).

□

Proof of Proposition 3.5.1

Proof. Given the product topology on $\mathcal{B}$ (Assumption 3.1.3) and the a.s. *continuous Hadamard differentiability* (CHD) of $\tilde{\boldsymbol{\beta}}^i_{T,S} : \Omega \times \Theta \to \mathcal{B}$ on $\Theta\ \forall (T,S,i) \in \mathbb{N} \times \mathbb{N} \times \mathbb{N}$ (part *(i)* of Assumption 3.5.4), it follows immediately by Proposition C.11 and Corollary C.12 that the empirical binding function $\tilde{\boldsymbol{\beta}}_{T,S} : \Omega \times \Theta \to \mathcal{B}$ is likewise a.s. CHD on $\Theta\ \forall (T,S) \in \mathbb{N} \times \mathbb{N}$. Trivial algebra shows that the same holds for the centered empirical binding function $\Delta_{T,S} : \Omega \times \Theta \to \mathcal{B}$. Finally, by the CHD of $\mu_T : \mathcal{B} \to \mathbb{R}$ on $\mathcal{B}\ \forall T \in \mathbb{N}$ (part (v) of Assumption 3.5.4), the chain rule (Lemma C.9), and the continuity of continuous compositions (Lemma A.29) we obtain the first desired result,

$$Q_{T,S} : \Omega \times \Theta \to \mathbb{R} \text{ is a.s. CHD on } \Theta\, \forall (T,S) \in \mathbb{N} \times \mathbb{N},$$

with derivative,

$$\nabla Q_{T,S}(\boldsymbol{\theta}, \cdot) = \nabla \mu_T \Big(\Delta_{T,S}(\boldsymbol{\theta}), \nabla \Delta_{T,S}(\boldsymbol{\theta}, \cdot) \Big)\ \forall (\boldsymbol{\theta}, T, S) \in \Theta \times \mathbb{N} \times \mathbb{N}.$$

We now turn to the first result involving the novel smoothness concept of *uniform Hadamard equi-differentiability of the third kind* (UHED3) (Definition C.24) introduced in Section C.2 of Appendix C.

Condition *(i)* of Assumption 3.5.4 together with Assumption 3.1.3 implies by Proposition C.40 that $\{\tilde{\boldsymbol{\beta}}_{T,S}\}_{T \in \mathbb{N}}$ is a.s. UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. This implies trivially that $\{\Delta_{T,S}\}_{T \in \mathbb{N}}$ is also a.s. UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$.

Condition *(ii)* of Assumption 3.5.4 together with Assumption 3.1.3 implies by Proposition C.40 that for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta$ the sequence, $\{\nabla \tilde{\boldsymbol{\beta}}_{T,S}(\cdot, \boldsymbol{\theta}'_T)\}_{T \in \mathbb{N}}$ is a.s. UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. This implies trivially that for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta$ the sequence, $\{\nabla \Delta_{T,S}(\cdot, \boldsymbol{\theta}'_T)\}_{T \in \mathbb{N}}$ is also a.s. UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. Together with the above result that $\{\Delta_{T,S}\}_{T \in \mathbb{N}}$ is a.s. UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$ we obtain that for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta$,

$$\Big\{ \Big(\Delta_{T,S}\ ,\ \Delta_{T,S}(\cdot, \boldsymbol{\theta}'_T) \}_{T \in \mathbb{N}} \Big) \Big\}_{T \in \mathbb{N}} \quad \text{is a.s. UHED3 along sequences } \ \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0.$$

Finally, condition *(vi)* of Assumption 3.5.4 and the fact that,

$$\nabla Q_{T,S}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \nabla \mu_T \Big(\Delta_{T,S}(\boldsymbol{\theta}), \nabla \Delta_{T,S}(\boldsymbol{\theta}, \boldsymbol{\theta}') \Big)$$

yields by Propositions C.38 and C.39 the second result of interest,[28]

$$\text{For every } \ \boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta \ \text{ the sequence, } \Big\{ \nabla Q_{T,S}(\cdot, \boldsymbol{\theta}'_T) \}_{T \in \mathbb{N}} \Big) \Big\}_{T \in \mathbb{N}}$$
$$\text{is a.s. UHED3 along sequences } \ \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0.$$

[28] Note that the remaining conditions in Propositions C.38 C.39 are trivially satisfied.

By Assumption 3.5.3 and Proposition C.40, the third desired result follows,

$$\text{For every } \boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta \text{ the sequence, } \Big\{\nabla Q_{T,S}(\cdot, \mathbb{S}_{\Theta_T})\}_{T\in\mathbb{N}}\Big\}_{T\in\mathbb{N}}$$
$$\text{is a.s. UHED3 along sequences } \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0.$$

Let us now turn to the limit functions. Again, by Proposition C.11 and Corollary C.12, the CHD of $\boldsymbol{\beta}_i^* : \Theta \to \mathcal{B}_i \, \forall i \in \mathbb{N}$ on Θ implies the same property for $\boldsymbol{\beta}^* : \Theta \to \mathcal{B}$ and, by trivial algebra, the same holds for $\Delta_\infty : \Theta \to \mathcal{B}$. By the CHD of $\mu_\infty : \mathcal{B} \to \mathbb{R}$ (part *(vii)* of Assumption 3.5.4), the chain rule (Lemma C.9) and continuity of continuous compositions (Lemma A.29) we then have that Q_∞ is differentiable on Θ with derivative at $\boldsymbol{\theta}$ in the direction of $\boldsymbol{\theta}'$ given by,

$$\nabla Q_\infty(\boldsymbol{\theta}, \boldsymbol{\theta}') = \nabla \mu_\infty\Big(\Delta_\infty(\boldsymbol{\theta}), \nabla \Delta_\infty(\boldsymbol{\theta}, \boldsymbol{\theta}')\Big) \tag{3.29}$$

Another result of Proposition 3.5.1 is thus obtained,

$$Q_\infty : \Theta \to \mathbb{R} \text{ is CHD on } \Theta.$$

Note that by Lemma C.16 and Remark C.20 the CHD of Q_∞ implies immediately that,

$$Q_\infty : \Theta \to \mathbb{R} \text{ is UHD3 along sequences } \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0.$$

An analogous differentiability result can be derived for $\nabla Q_\infty(\cdot, \boldsymbol{\theta}) : \Theta \to \mathbb{R}$ on $S(\boldsymbol{\theta}_0, \epsilon) \subseteq \Theta$ for some $\epsilon > 0$ and every $\boldsymbol{\theta} \in \text{lin}(\Theta)$ from the CHD of $\nabla \boldsymbol{\beta}_i^*(\cdot, \boldsymbol{\theta}) : \Theta \to \mathcal{B}_i$ on $S(\boldsymbol{\theta}_0, \epsilon) \subseteq \Theta$ for some $\epsilon > 0$ and every $(\boldsymbol{\theta}, i) \in \text{lin}(\Theta) \times \mathbb{N}$ (part (iv) of Assumption 3.5.4) and the CHD of $\nabla \mu_\infty : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ on $\mathcal{B} \times \mathcal{B}$ (part (viii) of Assumption 3.5.4).[29] In particular, given the product topology on $\mathcal{B}$ (Assumption 3.1.3), it follows immediately by Proposition C.11 and Corollary C.12 that $\nabla \boldsymbol{\beta}^*(\cdot, \boldsymbol{\theta}) : \Theta \to \mathcal{B}$ is CHD on Θ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$. The same holds for $\nabla \Delta_\infty(\cdot, \boldsymbol{\theta}) : \Theta \to \mathcal{B} \, \forall \boldsymbol{\theta} \in \text{lin}(\Theta)$ since $\nabla \Delta_\infty = -\nabla \boldsymbol{\beta}^*$ trivially on Θ. By the CHD of $\nabla \mu_\infty : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ (part (vii) of Assumption 3.5.4), Proposition C.13, and the continuity of continuous compositions (Lemma A.29) we obtain that,

$$\nabla Q_\infty(\cdot, \boldsymbol{\theta}) : \Theta \to \mathbb{R} \text{ is CHD on } S(\boldsymbol{\theta}_0, \epsilon) \text{ for every } \boldsymbol{\theta} \in \text{lin}(\Theta).$$

As a result, by Assumption 3.5.3, Proposition C.11 and Corollary C.12 we obtain the desired result,

$$\nabla Q_\infty(\cdot, \mathbb{S}_\Theta) : \Theta \to \mathbb{R}^{|\mathbb{S}_\Theta|} \text{ is CHD on } S(\boldsymbol{\theta}_0, \epsilon) \text{ for some } \epsilon > 0.$$

Finally, we obtain the last result of interest. Condition *(iv)* of Assumption 3.5.4 together with Assumption 3.1.3 and Proposition C.40 implies that for every

[29] Clearly, differentiability of $\nabla \mu_\infty$ is only required on $\Big(\Delta_\infty(S(\boldsymbol{\theta}_0, \epsilon)), \nabla \Delta_\infty(S(\boldsymbol{\theta}_0, \epsilon), \Theta)\Big)$. There is however little to be gained (in applications) in terms of generality with such a change.

$\boldsymbol{\theta}'_T \to \boldsymbol{\theta}' \in \Theta$, the sequence $\{\nabla\boldsymbol{\beta}^*(\cdot, \boldsymbol{\theta}'_T)\}_{T\in\mathbb{N}}$ is UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. This implies trivially that for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta}' \in \Theta$, the sequence $\{\nabla\Delta_\infty(\cdot, \boldsymbol{\theta}'_T)\}_{T\in\mathbb{N}}$ is also UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. Now, by condition *(iii)* of Assumption 3.5.4, Lemma C.16 and Remark C.20 it also follows that $\boldsymbol{\beta}^*_i$ is UHD3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0 \ \forall i \in \mathbb{I}$. By Assumption 3.1.3 and Proposition C.11 this implies that $\boldsymbol{\beta}^*$ is UHD3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$. The same applies trivially to Δ_∞. Since, the sequence $\{\Delta_\infty\}_{T\in\mathbb{N}}$ is trivially UHED3 along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$ we obtain immediately that for every $\boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta$

$$\Big\{\big(\Delta_\infty\,,\ \nabla\Delta_\infty(\cdot, \boldsymbol{\theta}'_T)\big)\Big\}_{T\in\mathbb{N}} \text{ is UHED3 along sequences } \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0.$$

Finally, since the CHD of $\nabla\mu_\infty : \mathcal{B}\times\mathcal{B} \to \mathbb{R}$ on $\mathcal{B}\times\mathcal{B}$ (condition *(viii)* in Assumption 3.5.4) implies by Lemma C.16 and Remark C.20 the UHD3 (and trivially the UHED3 of $\{\mu_\infty\}_{T\in\mathbb{N}}$) along every convergent sequence in $\mathcal{B}\times\mathcal{B}$, it follows by Proposition C.38 that,

$$\text{For every } \boldsymbol{\theta}'_T \to \boldsymbol{\theta} \in \Theta \text{ the sequence } \Big\{\nabla Q_\infty(\cdot, \boldsymbol{\theta}'_T)\Big\}_{T\in\mathbb{N}}$$
$$\text{is UHED3 along sequences } \boldsymbol{\theta}_T \to \boldsymbol{\theta}_0.$$

□

Proof of Proposition 3.5.2

Proof. Continuity of product inverse maps is ensured by continuity of each inverse component in the product topology. Assumption 3.5.5 *(i)* implies (by Lemma A.44, Corollary A.49 and Proposition B.15) the continuous invertibility of $\nabla\boldsymbol{\beta}^*(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathcal{B})$. This implies naturally the continuous invertibility of

$$\nabla\Delta_\infty(\boldsymbol{\theta}_0, \cdot) := -\nabla\boldsymbol{\beta}^*(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathcal{B}).$$

Likewise, Assumption 3.5.5 *(ii)* implies (by Definition B.14 and Proposition B.15) the continuous invertibility of $\nabla\boldsymbol{\beta}^*_{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathcal{B})$ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$ where $\boldsymbol{\beta}^*_{\nabla_{\boldsymbol{\theta}}} := \nabla\boldsymbol{\beta}^*(\cdot, \boldsymbol{\theta})\ \forall\boldsymbol{\theta} \in \text{lin}(\Theta)$. Define $\Delta^{\nabla_{\boldsymbol{\theta}}}_\infty := \nabla\Delta_\infty(\cdot, \boldsymbol{\theta})$ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$. Again, this implies naturally the continuous invertibility of

$$\nabla\Delta^{\nabla_{\boldsymbol{\theta}}}_\infty(\boldsymbol{\theta}_0, \cdot) = -\nabla\boldsymbol{\beta}^*_{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathcal{B}) \quad \text{for every} \quad \boldsymbol{\theta} \in \text{lin}(\Theta).$$

Existence of the derivative function $Q^{\nabla_{\boldsymbol{\theta}}}_\infty := \nabla Q_\infty(\cdot, \boldsymbol{\theta}) : \Theta \to \mathbb{R}$ and the Hadamard derivative $\nabla Q^{\nabla_{\boldsymbol{\theta}}}_\infty(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}\Big(\text{lin}(\Theta), \mathbb{R}\Big)$ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$ is ensured by Proposition 3.5.1 under Assumptions 3.1.1-3.1.3 and 3.5.1-3.5.4. Now, define the map $\boldsymbol{\beta}^{\boldsymbol{\theta}}_\nabla : \Theta \to \mathcal{B}\times\mathcal{B}$ and the derivative $\nabla\boldsymbol{\beta}^{\boldsymbol{\theta}}_\nabla(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\Theta, \mathcal{B}\times\mathcal{B})$ as,

$$\boldsymbol{\beta}^{\boldsymbol{\theta}}_\nabla(\cdot) := \Big(\Delta_\infty(\cdot)\,, \nabla\Delta_\infty(\cdot, \boldsymbol{\theta})\Big) \quad \text{and} \quad \nabla\boldsymbol{\beta}^{\boldsymbol{\theta}}_\nabla(\boldsymbol{\theta}_0, \cdot) := \Big(\nabla\Delta_\infty(\boldsymbol{\theta}_0, \cdot)\,, \nabla\Delta^{\nabla_{\boldsymbol{\theta}}}_\infty(\boldsymbol{\theta}_0, \cdot)\Big),$$

for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$. Clearly, for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$ it holds true that $Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0) = \nabla\mu_\infty(\boldsymbol{\beta}_\nabla^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0))$. Also, by defining $\mu_\infty^\nabla := \nabla\mu_\infty$ and making use of the Chain Rule (Lemma C.9), we have that,

$$\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot) = \nabla\mu_\infty^\nabla\Big(\boldsymbol{\beta}_\nabla^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\,,\, \nabla\boldsymbol{\beta}_\nabla^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \cdot)\Big) \quad \forall\, \boldsymbol{\theta} \in \text{lin}(\Theta).$$

By Lemma A.44, Corollary A.49 and Proposition B.15, a vector function is continuously invertible uniformly on a parameter if its components are. Hence, the continuous invertibility of $\nabla\Delta_\infty(\boldsymbol{\theta}_0, \cdot)$ and $\nabla\Delta_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot)$ imply the continuous invertibility of $\nabla\boldsymbol{\beta}_\nabla^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \cdot)$ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$. This implies trivially, the uniform continuous invertibility of $\Big(\boldsymbol{\beta}_\nabla^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\,,\, \nabla\boldsymbol{\beta}_\nabla^{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \cdot)\Big)$. Together with the uniform continuous invertibility of $\nabla\mu_\infty^\nabla$, we obtain by Propositions B.16 that the composition is itself continuously invertible for every $\boldsymbol{\theta}$ and hence obtain the desired result,

$$\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathbb{R}) \text{ is continuously invertible for every } \boldsymbol{\theta} \in \text{lin}(\Theta).$$

Finally, by appealing to Proposition B.15 it follows immediately under Assumption 3.5.3 that,

$$\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\text{lin}\Theta, \mathbb{R}^{|\mathbb{S}_\Theta|}) \text{ is continuously invertible.}$$

□

Proof of Proposition 3.5.3

Under Assumptions 3.1.1-3.1.3 and 3.5.1-3.5.4, Proposition 3.5.1 holds true. Hence, both $Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}') = \nabla Q_\infty(\boldsymbol{\theta}', \boldsymbol{\theta})$ and $\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ are well defined directional derivatives for every $(\boldsymbol{\theta}', \boldsymbol{\theta}) \in \Theta \times \text{lin}(\Theta)$. Under the added influence of Assumption 3.5.5, Proposition 3.5.2 obtained the continuous invertibility of $\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot)$ for every $\boldsymbol{\theta}$. Now, using $\boldsymbol{\theta}_T^0 \to \boldsymbol{\theta}_0$ (denseness Assumption 3.1.2 plus CHD of Q_∞ in Proposition 3.5.1), the continuity of $\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}_0, \cdot)$ on $\boldsymbol{\theta}$ (Definition C.3), the CMT (Corollary A.55) and Lemma A.79 and Remarks A.80 and A.81 for compact convergence of linear operators, it follows by Lemma B.13 and Proposition B.18 that,

$$\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta \leq \bar{c}\Big|\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\Big|$$

and immediately that,

$$\begin{aligned}
\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta &\leq \bar{c}\Big|\nabla Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\Big| \\
&\leq \bar{c}\Big|Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_T^0) - Q_\infty^{\nabla_{\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0}}(\boldsymbol{\theta}_0)\Big| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \\
&= \bar{c}\Big|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0) - \nabla Q_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\Big| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) \\
&= \bar{c}\Big|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0)\Big| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|)
\end{aligned} \tag{3.30}$$

where the second inequality follows by the *uniform Hadamard equi-differentiability* (Definition C.24) of $Q_\infty^{\nabla_{\boldsymbol{\theta}_T}}$ along sequences $\boldsymbol{\theta}_T \to \boldsymbol{\theta}_0$ for every $\boldsymbol{\theta}_T \to \boldsymbol{\theta}$ in Proposition 3.5.1 and Remark C.25. The first equality holds by definition and the second from having $\nabla Q_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = 0 \,\forall\, \boldsymbol{\theta} \in \Theta$ (Lemma C.10).[30] As a result, it follows that for large enough T,

$$\begin{aligned}\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta &\leq \bar{c}\Big|Q_\infty(\boldsymbol{\theta}_T^0) - Q_\infty(\boldsymbol{\theta}_0)\Big| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta)\\ &\leq \bar{c}\Big|Q_\infty(\pi_T\boldsymbol{\theta}_0) - Q_\infty(\boldsymbol{\theta}_0)\Big| + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta)\\ &\leq O(\|\pi_T\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0\|_\Theta) + o(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta)\end{aligned}$$

where the first inequality follows from (3.30) and by the CHD (and resulting UHD) of Q_∞ in Proposition 3.5.1, the second inequality is obtained since $Q_\infty(\pi_T\boldsymbol{\theta}_0) \geq Q_\infty(\boldsymbol{\theta}_T^0) \geq Q_\infty(\boldsymbol{\theta}_0)$ by construction, and the last inequality follows again simply by applying the definition of Hadamard differentiable operator. Finally, by Assumption 3.5.6, we obtain,

$$\begin{aligned}&\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta(1 + o(1)) = O(\|\pi_T\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0\|_\Theta)\\ &\Leftrightarrow \|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|_\Theta = \frac{1}{(1 + o(1))}O(o(T^{-1/2})) = o(T^{-1/2}).\end{aligned}$$

It now follows easily that $\left\|Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}$ converges to zero at an appropriate rate. In particular,

$$\begin{aligned}\left\|Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}} &= \left\|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}} = \left\|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) - \nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}\\ &= O(\|\boldsymbol{\theta}_T^0 - \boldsymbol{\theta}_0\|) = O\Big(o\Big(T^{-1/2}\Big)\Big) = o\Big(T^{-1/2}\Big).\end{aligned}$$

where the first equality follows by definition, the second follows from the fact that $\nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta) = 0$ (Lemma C.10), and the third by the continuous Hadamard differentiability of $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta)$ at $\boldsymbol{\theta}_0$ derived in Proposition 3.5.1 under the present set of assumptions. We thus state for future reference that,

$$\nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta) = 0 \quad \text{and} \quad \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) = o\Big(T^{-1/2}\Big). \tag{3.31}$$

Proof of Proposition 3.5.4

Proof. Note first that convergence to a tight Gaussian process $\mathbb{G}_S(\boldsymbol{\theta}_0)$,

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0^*) \xrightarrow{d} \mathbb{G}_S(\boldsymbol{\theta}_0),$$

[30] Note here that a differentiable function f satisfies $\|f(a) - f(a_0) - \nabla f(a_0, a - a_0)\| = o(\|a - a_0\|) \Leftrightarrow \|f(a) - f(a_0)\| + \|\nabla f(a_0, a - a_0)\| \geq o(\|a - a_0\|) \Leftrightarrow \|f(a) - f(a_0)\| + \|\nabla f(a_0, a - a_0)\| + o(\|a - a_0\|) \geq 0 \Leftrightarrow \|f(a) - f(a_0)\| + o(\|a - a_0\|) \geq -\|\nabla f(a_0, a - a_0)\| \Leftrightarrow \|f(a) - f(a_0)\| + o(\|a - a_0\|) \geq \|-\nabla f(a_0, a - a_0)\| \Leftrightarrow \|f(a) - f(a_0)\| + o(\|a - a_0\|) \geq \|\nabla f(a_0, a - a_0)\|$ by noting that $-o(\|a - a_0\|) = o(\|a - a_0\|)$ and that $\|-\nabla f(a_0, a - a_0)\| = |-1|\|\nabla f(a_0, a - a_0)\| = \|\nabla f(a_0, a - a_0)\|$.

follows immediately, under item *(i)* of Assumption 3.5.7 and separability of Θ (Assumption 3.1.1), by Definitions A.72 and A.73 and Lemmas A.74, A.75 and A.61. Equivalently,

$$\sqrt{T}(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) - \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)) \xrightarrow{d} \tilde{\mathbb{G}}_S(\boldsymbol{\theta}_0),$$

follows from point *(ii)* of Assumption 3.5.7 by the same argument. This implies that the sequences,

$$\left\{\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0^*)\right\}_{t\in\mathbb{N}} \quad \text{and} \quad \left\{\sqrt{T}(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\right\}_{T\in\mathbb{N}}$$

are individually asymptotically tight (See Definition A.73 and Lemma A.74). By Lemma A.76,

$$\left\{\left(\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0^*)\,,\ \sqrt{T}(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\right)\right\}_{T\in\mathbb{N}}$$

is also asymptotically tight. This implies by the continuity of vector addition in topological vector spaces (Lemma A.28), the tightness of continuous transformations and the *Continuous Mapping Theorem* (Corollary A.55) that,

$$\begin{aligned}\sqrt{T}\Big(\big(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0^*\big) - \big(\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0) - \boldsymbol{\beta}^*(\boldsymbol{\theta}_0)\big)\Big) &= \sqrt{T}\big(\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0)\big) \\ &\xrightarrow{d} \mathbb{G}_S(\boldsymbol{\theta}_0) - \tilde{\mathbb{G}}_S(\boldsymbol{\theta}_0) := \mathbb{G}_\Delta^S(\boldsymbol{\theta}_0)\end{aligned}$$

where $\mathbb{G}_\Delta^S(\boldsymbol{\theta}_0)$ is again a tight Gaussian process.

Now, recall that both $\nabla\tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T})$ and $\nabla\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)$ take values in $\mathcal{B}_i^{|\mathbb{S}_\Theta|}$. Given the product topology on the product space $\mathcal{B}^{|\mathbb{S}_\Theta|}$ (Assumption 3.5.3) we obtain naturally that point *(iii)* of Assumption 3.5.7 implies by Lemma A.61,

$$\begin{aligned}&\|\nabla_{\Theta_T}\tilde{\boldsymbol{\beta}}_{T,S}^i(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla_\Theta\boldsymbol{\beta}_i^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\|_{\mathcal{B}} = o_p(r_T)\ \forall\,(i,S) \in \mathbb{N}\times\mathbb{N} \\ &\Rightarrow\ \|\nabla_{\Theta_T}\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla_\Theta\boldsymbol{\beta}^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\|_{\mathcal{B}} = o_p(r_T)\ \forall\, S \in \mathbb{N}.\end{aligned}$$

Together with differentiability of μ_∞ (Assumption 3.5.4), the uniform convergence in point *(iv)* of Assumption 3.5.7 $\sup_{(\boldsymbol{\beta},\boldsymbol{\beta}')\in\mathcal{B}\times\mathcal{B}}\left|\nabla\mu_T(\boldsymbol{\beta},\boldsymbol{\beta}') - \nabla\mu_\infty(\boldsymbol{\beta},\boldsymbol{\beta}')\right| = o(T^{-1/2})$ as $T\to\infty$ and the order-of-magnitude on the first argument in point *(v)* of Assumption 3.5.7 $\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left|\nabla\mu_\infty(\boldsymbol{\beta},\boldsymbol{\beta}_T)\right| = o\big(\xi_\mu(\|\boldsymbol{\beta}_T\|_{\mathcal{B}})\big)$ for every $\boldsymbol{\beta}_T \to 0$ where $r_T = \xi_\mu^{-1}(T^{-1/2})$, we obtain by Proposition C.32 that,

$$\left\{\nabla\mu_T(\cdot, \nabla\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}))\right\}_{T\in\mathbb{N}}$$

is a $\sqrt{T}$-*Hadamard Sequence* w.r.t. $\mu_\infty(\cdot, \nabla\boldsymbol{\beta}^*(\boldsymbol{\theta}_0, \mathbb{S}_\Theta))$ at the origin of $\mathbb{B}$. Finally, Proposition C.34 implies the desired result.

□

Proof of Proposition 3.5.5

Proof. Note first that by norm sub-additivity and linearity of derivatives,

$$\begin{aligned}
&\sqrt{T}\left\|\left(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}-Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}\right)(\hat{\boldsymbol{\theta}}_{T,S})-\left(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}-Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}\right)(\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}\\
&\quad=\sqrt{T}\left\|Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\hat{\boldsymbol{\theta}}_{T,S})-Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0)-Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\hat{\boldsymbol{\theta}}_{T,S})+Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}\\
&\quad\le\sqrt{T}\left\|Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\hat{\boldsymbol{\theta}}_{T,S})-Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0)-\nabla Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}\\
&\qquad+\left\|\nabla Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0))-\nabla Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0))\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}\\
&\qquad+\sqrt{T}\left\|Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\hat{\boldsymbol{\theta}}_{T,S})+Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0)-\nabla Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}.
\end{aligned}\tag{3.32}$$

The desired result will thus follow by having,

$$\sqrt{T}\left\|Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\hat{\boldsymbol{\theta}}_{T,S})+Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0)-\nabla Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}=o(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta),\tag{3.33}$$

$$\left\|\nabla Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0))-\nabla Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0))\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}=o(1),\tag{3.34}$$

and

$$\sqrt{T}\left\|Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\hat{\boldsymbol{\theta}}_{T,S})-Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0)-\nabla Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}=o(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta).\tag{3.35}$$

Indeed, conditions (3.33), (3.34) and (3.35) imply by (3.32) that,

$$\begin{aligned}
&\sqrt{T}\left\|\left(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}-Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}\right)(\hat{\boldsymbol{\theta}}_{T,S})-\left(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}-Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}\right)(\boldsymbol{\theta}_T^0)\right\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}}\\
&=\sqrt{T}o_p(\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|)+o_p(1)+\sqrt{T}o(\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|)=o_p(1+\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|).
\end{aligned}$$

Condition (3.33) follows by the continuous Hadamard differentiability of $Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}:\Theta\to\mathbb{R}^{|\mathbb{S}_\Theta|}$ in a neighborhood of $\boldsymbol{\theta}_0$ derived in Proposition 3.5.1, and consequent uniform Hadamard differentiability of the third kind (Definition C.18 and Remark C.19) along sequences $\boldsymbol{\theta}_T^0\to\boldsymbol{\theta}_0$.

Condition (3.35) follows by the a.s. UHED3 (Definition C.24 and Remark C.25) of $Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}$ at $\boldsymbol{\theta}_T^0$ for every $T\in\mathbb{N}$ established in Proposition 3.5.1.

Condition (3.34) is obtained as follows. Notice that for every $\boldsymbol{\theta}\in\mathrm{lin}(\Theta)$,

$$\nabla Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}_T^0,\boldsymbol{\theta})=\nabla\mu_T^{\nabla}\left(\boldsymbol{\beta}_{\nabla_{T,S}}^{\mathbb{S}_{\Theta_T}}(\boldsymbol{\theta}_T^0)\,,\,\nabla\boldsymbol{\beta}_{\nabla_{T,S}}^{\mathbb{S}_{\Theta_T}}(\boldsymbol{\theta}_T^0,\boldsymbol{\theta})\right)$$

$$\text{and}\quad\nabla Q_{\infty}^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0,\boldsymbol{\theta})=\nabla\mu_{\infty}^{\nabla}\left(\boldsymbol{\beta}_{\nabla}^{\mathbb{S}_\Theta}(\boldsymbol{\theta}_T^0)\,,\,\nabla\boldsymbol{\beta}_{\nabla}^{\mathbb{S}_\Theta}(\boldsymbol{\theta}_T^0,\boldsymbol{\theta})\right)$$

$$\text{where}\quad\boldsymbol{\beta}_{\nabla_{T,S}}^{\mathbb{S}_{\Theta_T}}(\boldsymbol{\theta}_T^0):=\left(\Delta_{T,S}(\boldsymbol{\theta}_T^0)\,,\,\nabla\Delta_{T,S}(\boldsymbol{\theta}_T^0,\mathbb{S}_{\Theta_T})\right),$$

$$\boldsymbol{\beta}_{\nabla}^{\mathbb{S}_\Theta}(\boldsymbol{\theta}_T^0):=\left(\Delta_{\infty}(\boldsymbol{\theta}_T^0)\,,\,\nabla\Delta_{\infty}(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\right),$$

$$\nabla\boldsymbol{\beta}^{\mathbb{S}_{\Theta_T}}_{\nabla_{T,S}}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) := \Big(\nabla\Delta_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T}),\, \nabla\Delta^{\nabla_{\mathbb{S}_{\Theta_T}}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})\Big),$$

$$\text{and } \nabla\boldsymbol{\beta}^{\mathbb{S}_\Theta}_{\nabla}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) := \Big(\nabla\Delta_{\infty}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta}),\, \nabla\Delta^{\nabla_{\mathbb{S}_\Theta}}_{\infty}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})\Big).$$

By Assumption 3.4.2, $\tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta}^0_T) \xrightarrow{p} \boldsymbol{\beta}^*_i(\boldsymbol{\theta}^0_T)$ holds for every $i \in \mathbb{N}$. This implies by Assumption 3.1.3 and Corollary A.16 that, $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^0_T) \xrightarrow{p} \boldsymbol{\beta}^*(\boldsymbol{\theta}^0_T)$. Together with $\hat{\boldsymbol{\beta}}_T \to \boldsymbol{\beta}^*_0$ this implies $\Delta_{T,S}(\boldsymbol{\theta}^0_T) \to \Delta_\infty(\boldsymbol{\theta}^0_T)$. By Assumption 3.5.7, $\nabla\tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) \xrightarrow{p} \nabla\boldsymbol{\beta}^*_i(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})$ holds for every $(i,\boldsymbol{\theta}) \in \mathbb{N}\times\mathbb{S}_\Theta$. This implies by Assumption 3.1.3 and Corollary A.16 that, $\nabla\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) \xrightarrow{p} \nabla\boldsymbol{\beta}^*(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})$ holds $\forall\boldsymbol{\theta} \in \mathbb{S}_\Theta$. Hence, by Assumption 3.5.3 and Corollary A.16, $\nabla\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T}) \xrightarrow{p} \nabla\boldsymbol{\beta}^*(\boldsymbol{\theta}^0_T,\mathbb{S}_\Theta)$. Now, since $\nabla\Delta_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T}) := -\nabla\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T})$ and $\nabla\Delta_\infty(\boldsymbol{\theta}^0_T,\mathbb{S}_\Theta) := -\nabla\boldsymbol{\beta}^*(\boldsymbol{\theta}^0_T,\mathbb{S}_T)$ we obtain $\nabla\Delta_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T}) \to \nabla\Delta_\infty(\boldsymbol{\theta}^0_T,\mathbb{S}_\Theta)$. Finally, by Assumption 3.5.8, we have that $\nabla\tilde{\boldsymbol{\beta}}^{\nabla_{\boldsymbol{\theta}},i}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}') \xrightarrow{p} \nabla\boldsymbol{\beta}^{*,i}_{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}')$ holds for every $(i,\boldsymbol{\theta},\boldsymbol{\theta}') \in \mathbb{N}\times\mathbb{S}_\Theta\times\mathrm{lin}\Theta$. This implies by Assumption 3.1.3 and Corollary A.16 that, $\nabla\tilde{\boldsymbol{\beta}}^{\nabla_{\boldsymbol{\theta}}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}') \xrightarrow{p} \nabla\boldsymbol{\beta}^*_{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}')$ holds $\forall(\boldsymbol{\theta},\boldsymbol{\theta}') \in \mathbb{S}_\Theta\times\mathrm{lin}\Theta$. By Assumption 3.5.3 and Corollary A.16 we then have $\nabla\tilde{\boldsymbol{\beta}}^{\nabla_{\mathbb{S}_{\Theta_T}}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}') \xrightarrow{p} \nabla\boldsymbol{\beta}^*_{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}')\,\forall\boldsymbol{\theta}' \in \mathrm{lin}\Theta$. Now, since we have $\nabla\Delta^{\nabla_{\mathbb{S}}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) = -\nabla\tilde{\boldsymbol{\beta}}^{T,S}_{\nabla_{\mathbb{S}_{\Theta_T}}}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})$ and $\nabla\Delta^{\nabla_{\mathbb{S}_\Theta}}_{\infty}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) = -\nabla\boldsymbol{\beta}^*_{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})$, we have that $\nabla\Delta^{\nabla_{\mathbb{S}}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta}) \xrightarrow{p} \nabla\Delta^{\nabla_{\mathbb{S}_\Theta}}_{\infty}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})\,\forall\boldsymbol{\theta} \in \mathrm{lin}\Theta$. Furthermore, since $\nabla\Delta^{\nabla_{\mathbb{S}}}_{T,S}(\boldsymbol{\theta}^0_T,\cdot)$ is a bounded linear operator for every $T \in \mathbb{N}$, we have that the convergence is uniform on compact sets by Lemma A.79, Remarks A.80 and A.81 and its stochastic counterpart in Lemma 1.1.2 discussed in Section 1.1. We thus obtain,

$$\Big(\Delta_{T,S}(\boldsymbol{\theta}^0_T),\, \nabla\Delta_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T})\Big) \xrightarrow{p} \Big(\Delta_\infty(\boldsymbol{\theta}^0_T),\, \nabla\Delta_\infty(\boldsymbol{\theta}^0_T,\mathbb{S}_\Theta)\Big),$$

and

$$\Big(\nabla\Delta_{T,S}(\boldsymbol{\theta}^0_T,\mathbb{S}_{\Theta_T}),\, \nabla\Delta^{\nabla_{\mathbb{S}_{\Theta_T}}}_{T,S}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})\Big) \xrightarrow{p} \Big(\nabla\Delta_\infty(\boldsymbol{\theta}^0_T,\mathbb{S}_\Theta),\, \nabla\Delta^{\nabla_{\mathbb{S}_\Theta}}_{\infty}(\boldsymbol{\theta}^0_T,\boldsymbol{\theta})\Big)$$

uniformly in $\boldsymbol{\theta} \in \mathbb{A}^*$ for every compact $\mathbb{A}^* \subset \mathrm{lin}(\Theta)$. Together with the uniform convergence of $\nabla\mu^\nabla_T \to \nabla\mu^\nabla_\infty$ (Assumption 3.5.7) on compact sets, this implies,

$$\sup_{\boldsymbol{\theta}\in\Theta^*}\Big\|\nabla Q^{\nabla_{\mathbb{S}_{\Theta_T}}}_{T,S}(\boldsymbol{\theta}^0_T,\cdot) - \nabla Q^{\nabla_{\mathbb{S}_\Theta}}_{\infty}(\boldsymbol{\theta}^0_T,\cdot)\Big\|_{\mathbb{R}^{|\mathbb{S}_\Theta|}} = o_p(1) \text{ for every compact } \Theta^* \subset \Theta.$$

The desired result in thus obtained by combining compact convergence with the tightness of the sequence $\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^0_T)$ on the separable set $\mathrm{lin}(\Theta)$.

□

Proof of Theorem 3.5.1

Proof. Note first that by Theorem 3.4.1, Assumptions 3.1.1-3.4.5 imply that,

(i) $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0\| \xrightarrow{p} 0.$

The assumption that $\hat{\boldsymbol{\theta}}_{T,S}$ is an exact SNPII estimator as in (3.2) implies that $\hat{\boldsymbol{\theta}}_{T,S} = \boldsymbol{\theta}^*_{T,S}$ in (2.19). The assumption that $\{\Theta_T\}_{T\in\mathbb{N}}$ are purely dimensional sieves w.r.t. the sequence $\{Q_{T,S}\}_{T\in\mathbb{N}}$ implies that $\boldsymbol{\theta}^*_{T,S} = \boldsymbol{\theta}^{**}_{T,S}$ in (2.19). We thus have $\hat{\boldsymbol{\theta}}_{T,S} = \boldsymbol{\theta}^{**}_{T,S}$. Now, together together with the Hadamard differentiability of $Q_{T,S}\,\forall T \in \mathbb{N}$ (obtained in Proposition 3.5.1 under Assumption 3.5.4, this implies by the generalized Fermat's stationary points theorem (Lemma C.10) that

(ii) $\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_{\Theta_T}) = 0\,\forall T \in \mathbb{N}$.

Also, Propositions 3.5.1, 3.5.2, 3.5.3, 3.5.4 and 3.5.5 establish together that,

(iii) $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta) : \Theta \to \mathbb{R}$ is CHD on a neighborhood of $\boldsymbol{\theta}_0$;

(iv) $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot) \in \mathbb{L}(\Theta, \mathbb{R}^{|\mathbb{S}_\Theta|})$ is continuously invertible;

(v) $\|\boldsymbol{\theta}^0_T - \boldsymbol{\theta}_0\|_\Theta = o(T^{-1/2})$ and $\nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}) = 0$;

(vi) $\sqrt{T}\Big[\nabla Q_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\Big] \xrightarrow{d} \mathbb{G}_S(\boldsymbol{\theta}_0)$ as $T \to \infty$ where $\mathbb{G}_S(\boldsymbol{\theta}_0)$ is a tight Gaussian process;

(vii) $\sqrt{T}\Big[\Big(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}} - Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\Big)(\hat{\boldsymbol{\theta}}_{T,S}) - \Big(Q_{T,S}^{\nabla_{\mathbb{S}_{\Theta_T}}} - Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\Big)(\boldsymbol{\theta}^0_T)\Big] = o_p(1+\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^0_T\|_\Theta)$.

Finally, the desired result is obtained from the items *(i)-(vi)* by an adaptation of the convergence theorem in van der Vaart (1995) and Theorem 3.3.1 in van der Vaart and Wellner (1996) (Lemma A.90) as presented in Theorem 2.3.1 in Chapter 2 for the general case of sieve extremum estimators under high-level assumptions.

By Proposition 3.5.3 there exists $T^* \in \mathbb{N}$ such that $\boldsymbol{\theta}^0_T \in S_{\boldsymbol{\theta}_0}(\epsilon)\,\forall T \geq T^*$. Proposition 3.5.1 ensures that $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}^0_T, \cdot) \in \mathbb{L}(\mathrm{lin}(\Theta), \mathbb{R}^{|\mathbb{S}_\Theta|})$ is defined for every $T > T^*$. By Proposition 3.5.2, $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot)$ has a continuous inverse defined on its range. Finally, by Lemma B.13 it holds true that,

$$\exists\, \bar{c} \in \mathbb{R}^+_0 \text{ such that } \Big\|\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \boldsymbol{\theta} - \boldsymbol{\theta}_0)\Big\| \geq \bar{c}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\Theta \quad \forall \boldsymbol{\theta} \,:\, (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \mathrm{lin}(\Theta).$$

Now, by continuous differentiability and the compact convergence of bounded linear maps (Lemma A.79, Remarks A.80 and A.81) we obtain,

$$\Big|\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}^0_T, \boldsymbol{\theta}_T) - \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})\Big| \to 0$$

as $T \to \infty$ for every sequence $\boldsymbol{\theta}_T \to \boldsymbol{\theta} \in \Theta$ with $\boldsymbol{\theta}_T \in \mathrm{lin}(\Theta_T)\,\forall T > T*$. This in turn implies by Proposition B.18 that $\exists\, c > 0$ such that,

$$\Big\|\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}^0_T, \boldsymbol{\theta}_T - \boldsymbol{\theta}^0_T)\Big\| \geq c\|\boldsymbol{\theta}_T - \boldsymbol{\theta}^0_T\|_\Theta \tag{3.36}$$

holds for every sequence $\{\boldsymbol{\theta}_T - \boldsymbol{\theta}^0_T\}_{T\in\mathbb{N}}$ such that $(\boldsymbol{\theta}_T - \boldsymbol{\theta}^0_T) \in \mathrm{lin}(\Theta_T)\,\forall T > T^*$.

Now, the continuous Hadamard differentiability of $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta) : \Theta \to \mathbb{R}^{|\mathbb{S}_\Theta|}$ on $S_{\boldsymbol{\theta}_0}(\epsilon)$ postulated in Proposition 3.5.1 implies also by Lemma C.16 and Remark C.19 that $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta)$ is uniformly Hadamard differentiable of the third kind along every sequence $\boldsymbol{\theta}_T^0 \to \boldsymbol{\theta}_0$. In particular,

$$\left\| \nabla Q_\infty(\boldsymbol{\theta}_T, \mathbb{S}_\Theta) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) - \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0, \boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0) \right\| = o(\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0\|_\Theta) \quad (3.37)$$

holds for every sequence $\{\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0\}_{T\in\mathbb{N}}$ such that $(\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0) \in \text{lin}(\Theta_T)\, \forall T > T^*$. Hence, using (3.36), it follows immediately that,

$$\begin{aligned} &\left\| \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_T^0, \boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0) \right\| \geq c\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0\|_\Theta \\ &\Leftrightarrow \left\| \nabla Q_\infty(\boldsymbol{\theta}_T, \mathbb{S}_\Theta) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) \right\| \geq c\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0\|_\Theta + o(\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0\|_\Theta) \end{aligned}$$

holds also for every sequence $\{\boldsymbol{\theta}_T - \boldsymbol{\theta}_T^0\}_{T\in\mathbb{N}}$ in $\text{lin}(\Theta_T)\, \forall T > T^*$. Finally, we can conclude that,

$$\begin{aligned} \sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta \Big(c + o(1)\Big) &\leq \sqrt{T} \left\| \nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_\Theta) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) \right\| \\ &\leq \sqrt{T} \left\| \nabla Q_{T,S}(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) \right\| \\ &\quad + o_p(1 + \sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta) \end{aligned} \quad (3.38)$$

holds again for every sequence $\{\hat{\boldsymbol{\theta}}_{T,S}\}_{T\in\mathbb{N}}$ and $\{\boldsymbol{\theta}_T^0\}_{T\in\mathbb{N}}$ such that $(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0) \in \Theta_T\, \forall T \in \mathbb{N}$ where the first inequality in (3.38) is obtained from the last inequality (3.11) simply by multiplying both sides by $\sqrt{T}$ and rewriting $c\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta - o(\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta)$ as $\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta \Big(c + o(1)\Big)$. The second follows by noting that,

$$\begin{aligned} \sqrt{T} \left\| \nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_\Theta) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) \right\| &\leq \sqrt{T} \left\| \nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_\Theta) \right\| + o(1) \\ &\leq \sqrt{T} \Big\| \nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_\Theta) - \\ &\qquad \nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_{\Theta_T}) \Big\| + o_p(1) \\ &= -\sqrt{T} \left\| \nabla Q_{T,S}(\boldsymbol{\theta}_T^0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) \right\| \\ &\qquad + o_p(1 + \sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta) \end{aligned} \quad (3.39)$$

where the first inequality in (3.39) holds since $\sqrt{T}\|\nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta)\| = o(T^{-1/2})$ (Proposition 3.5.3) the second by adding and subtracting $\sqrt{T}\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_{\Theta_T})$, norm sub-additivity and noting that the condition $\sqrt{T}\|\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_{\Theta_T})\| = o_p(1)$ holds trivially. The last step follows immediately from Proposition 3.5.5. Finally

the desired result follows from (3.38) by noting that,

$$\begin{aligned}\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta\Big(c+o(1)\Big) &\leq \sqrt{T}\Big\|\nabla Q_{T,S}(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\Big\| \\ &\quad+\sqrt{T}\Big\|\nabla Q_{T,S}(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)\Big\| \\ &\quad+o_p(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta) \\ &\leq \sqrt{T}\Big\|\nabla Q_{T,S}(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\Big\| \\ &\quad+o(1)+o_p(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta)\end{aligned}$$

by adding and subtracting $\nabla Q_\infty(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)$ and noting that $\nabla Q_\infty(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)=0$ and thus concluding,

$$\begin{aligned}&\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta\Big(c+o(1)-o_p(1)\Big)\leq O_p(1) \\ \Leftrightarrow\quad &\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta\leq\frac{O_p(1)}{c+o_p(1)}=O_p(1)\end{aligned}$$

from the fact that $\sqrt{T}\Big\|\nabla Q_{T,S}(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\Big\|=O_p(1)$ (Proposition 3.5.4) which in turn implies naturally the $\sqrt{T}$ convergence of the SNPII estimator,

$$\begin{aligned}\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0\|_\Theta &= \sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0+\boldsymbol{\theta}_T^0-\boldsymbol{\theta}_0\|_\Theta \\ &\leq \sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta+\sqrt{T}\|\boldsymbol{\theta}_T^0-\boldsymbol{\theta}_0\|_\Theta \\ &= O_p(1)+o_p(1)=O_p(1).\end{aligned}$$

Asymptotic normality follows from having,

$$\begin{aligned}\sqrt{T}\Big[\nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\Big] &= \sqrt{T}\Big[\nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_\Theta)\Big]+o(1) \\ &= \sqrt{T}\Big[\nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_\Theta) \\ &\qquad-\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_{\Theta_T})\Big]+o_p(1) \\ &= -\sqrt{T}\Big[\nabla Q_{T,S}(\boldsymbol{\theta}_T^0,\mathbb{S}_{\Theta_T})-\nabla Q_\infty(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\Big] \\ &\qquad+o_p(1+\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_T^0\|_\Theta)\end{aligned}$$

which holds essentially by the same argument as in (3.39). This in turn implies that,

$$\begin{aligned}\sqrt{T}\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0,\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0) &= \sqrt{T}\Big[\nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_0,\mathbb{S}_\Theta)\Big] \\ &\quad+o_p(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0)\|) \\ &= \sqrt{T}\Big[\nabla Q_\infty(\hat{\boldsymbol{\theta}}_{T,S},\mathbb{S}_\Theta)-\nabla Q_\infty(\boldsymbol{\theta}_T^0,\mathbb{S}_\Theta)\Big] \\ &\quad+o_p(1)+o_p(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}_0)\|)\end{aligned} \tag{3.40}$$

by the Hadamard differentiability of $\nabla Q_\infty(\cdot, \mathbb{S}_\Theta)$ at $\boldsymbol{\theta}_0$ (Proposition 3.5.1) and $\nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta) = o(T^{-1/2})$ (Proposition 3.5.3). This implies naturally that,

$$\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\Big(\boldsymbol{\theta}_0, \sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0)\Big) = -\sqrt{T}\Big[\nabla Q_{T,S}(\boldsymbol{\theta}_T^0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta)\Big] + o_p(1 + \sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_T^0\|_\Theta) + o_p(\sqrt{T}\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0)\|) \tag{3.41}$$

and equivalently, by the continuous invertibility of $\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}$ at $\boldsymbol{\theta}_0$ (Proposition 3.5.2),

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0)\Big) = -\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\Big(\boldsymbol{\theta}_0, \cdot\Big)\Big)\Big(\sqrt{T}\Big[\nabla Q_{T,S}(\boldsymbol{\theta}_T^0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta)\Big]\Big) + o_p(1). \tag{3.42}$$

Finally, by noting that,

$$\Big[\nabla Q_{T,S}(\boldsymbol{\theta}_T^0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_T^0, \mathbb{S}_\Theta)\Big] = \Big[\nabla Q_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla Q_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)\Big] + o_p(T^{-1/2})$$

follows easily by the maintained uniform Hadamard equi-differentiability conditions, the desired result follows by the weak convergence in condition *(vi)* above,

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}_0)\Big) \xrightarrow{d} -\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}\Big(\boldsymbol{\theta}_0, \cdot\Big)\Big)(\mathbb{G}_0).$$

□

Proof of Theorem 3.5.2

Proof. Note first that all conditions *(i)-(viii)* laid down in the proof of Theorem 3.5.1 except for condition *(ii)*. We can however derive an alternative condition

(ii') $\nabla Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \mathbb{S}_{\Theta_T}) = o_p(T^{-1/2})\ \forall T \in \mathbb{N}$.

First, under Assumptions 3.1.1-3.1.3 and 3.5.1-3.5.5, Propositions 3.5.1 and 3.5.2 hold true. Hence, both $Q_{T,S}^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}') = \nabla Q_{T,S}(\boldsymbol{\theta}', \boldsymbol{\theta})$ and $\nabla Q_{T,S}^{\nabla_{\boldsymbol{\theta}}}(\boldsymbol{\theta}'', \boldsymbol{\theta})$ are well defined directional derivatives for every $(\boldsymbol{\theta}'', \boldsymbol{\theta}', \boldsymbol{\theta}) \in S(\boldsymbol{\theta}_0, \epsilon) \times \Theta \times \text{lin}(\Theta)$. By Corollary 3.4.1, under Assumptions 3.1.1-3.4.5, it follows immediately that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_\Theta = o_p(1)$ and hence that $\forall \epsilon > 0$, $\exists\, T^* \in \mathbb{N}$ such that $\boldsymbol{\theta}^* \in S(\boldsymbol{\theta}_0, \epsilon)$ with probability tending to one. By the invertibility conditions derived in Proposition 3.5.2, the pointwise convergence of derivatives in Assumptions 3.5.7 and 3.5.9 and the resulting compact convergence of linear operators (Lemma A.79 and 1.1.2 and Remarks A.80 and A.81) and Proposition B.18 we have that for every $T > T^*$ we thus have that $\exists\, \bar{c} > 0$ such

that,

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta &\leq \bar{c}\left|\nabla_{\Theta_T} Q_{T,S}^{\nabla_{\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}^*_{T,S}}}(\boldsymbol{\theta}^*_{T,S}, \hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S})\right| \\
&\leq \bar{c}\left|Q_{T,S}^{\nabla_{\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}^*_{T,S}}}(\hat{\boldsymbol{\theta}}_{T,S}) - Q_{T,S}^{\nabla_{\hat{\boldsymbol{\theta}}_{T,S}-\boldsymbol{\theta}^*_{T,S}}}(\boldsymbol{\theta}^*_{T,S})\right| + o(\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|) \\
&= \bar{c}\left|\nabla_{\Theta_T} Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}) - \nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^*_{T,S}, \hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S})\right| \\
&\quad + o(\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|) \\
&\leq \left|Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}) - Q_{T,S}(\boldsymbol{\theta}^*_{T,S})\right| + o(\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta) \\
&\qquad + \left|Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}) - Q_{T,S}(\boldsymbol{\theta}^*_{T,S})\right| + o(\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|)
\end{aligned}$$

where the second inequality follows by norm sub-additivity and the third using the UHED3 of $Q_{T,S}^{\nabla_{\boldsymbol{\theta}_T}}$ derived in Proposition 3.5.1. This implies since $\eta_T = \left|Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}) - Q_{T,S}(\boldsymbol{\theta}^*_{T,S})\right| = O_p(T^{-1/2})$ that,

$$\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta(1 + o(1)) \leq 2\eta_T \Leftrightarrow \|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta \leq \frac{2O_p(T^{-1/2})}{1 + o(1)} = O_p(T^{-1/2}).$$

Finally, the desired result follows by noting that,

$$\begin{aligned}
\left|\nabla_{\Theta_T} Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta})\right| &= \left|\nabla_{\Theta_T} Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}) - \nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^{**}_{T,S}, \boldsymbol{\theta})\right| \\
&\leq \left|\nabla_{\Theta_T} Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}) - \nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^*_{T,S}, \boldsymbol{\theta})\right| \\
&\qquad + \left|\nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^*_{T,S}, \boldsymbol{\theta}) - \nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^{**}_{T,S}, \boldsymbol{\theta})\right| \\
&= o(\|\hat{\boldsymbol{\theta}}_{T,S} - \boldsymbol{\theta}^*_{T,S}\|_\Theta) + o(\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta) \; \forall \boldsymbol{\theta} \in \Theta_T,
\end{aligned}$$

where the first equality follows by adding and subtracting $\nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^{**}_{T,S}, \boldsymbol{\theta}) = 0$, the inequality follows by adding and subtracting $\nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^*_{T,S}, \boldsymbol{\theta})$ and by norm sub-additivity, and the final step follows by the appropriate UHED3 of $\nabla_{\Theta_T} Q_{T,S}$ derived in Proposition 3.5.1. Now since, making use once more of the invertibility conditions derived in Proposition 3.5.2, the pointwise convergence of derivatives in Assumptions 3.5.7 and 3.5.9 and the resulting compact convergence of linear operators (Lemma A.79 and 1.1.2 and Remarks A.80 and A.81) and Proposition B.18 we have that for every $T > T^*$ we thus have that $\exists\, \bar{c} > 0$ such that,

$$\begin{aligned}
\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta &\leq \bar{c}\left|\nabla_{\Theta_T} Q_{T,S}(\boldsymbol{\theta}^*_{T,S}, \boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S})\right| \\
&\leq \bar{c}\left|Q_{T,S}(\boldsymbol{\theta}^*_{T,S}) - Q_{T,S}(\boldsymbol{\theta}^{**}_{T,S})\right| + o(\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta) \\
&\leq \bar{c}\left|Q_{T,S}(\pi_T(\boldsymbol{\theta}^{**}_{T,S})) - Q_{T,S}(\boldsymbol{\theta}^{**}_{T,S})\right| + o(\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta) \\
&\leq o(\|\pi_T(\boldsymbol{\theta}^{**}_{T,S}) - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta) + o(\|\boldsymbol{\theta}^*_{T,S} - \boldsymbol{\theta}^{**}_{T,S}\|_\Theta)
\end{aligned}$$

where the second inequality by applying the UHED3 property of $Q_{T,S}$, the third inequality is obtained since $Q_{T,S}(\pi_T(\boldsymbol{\theta}^{**}_{T,S})) \geq Q_{T,S}(\boldsymbol{\theta}^{*}_{T,S})$ and the last inequality by differentiability of $Q_{T,S}$. It thus follows by (3.11) that

$$\nabla_{\Theta_T} Q_{T,S}(\hat{\boldsymbol{\theta}}_{T,S}, \boldsymbol{\theta}) = o_p(T^{-1/2}) \, \forall \boldsymbol{\theta} \in \Theta_T. \tag{3.43}$$

condition. The desired result thus follows.

□

Proof of Theorem 3.6.1

Proof. Note first that $\pi_k\Big(\mathbb{G}_S(\boldsymbol{\theta}_0)\Big)$ is a probability measure on the Borel sigma-algebra $\mathfrak{B}(\mathcal{B})$ of $(\mathcal{B}, \mathcal{T}_{\mathcal{B}})$. Since $\mathcal{T}_{\mathcal{B}}$ is the product topology and $\mathfrak{B}(\mathcal{B})$ the product sigma-algebra (Assumption 3.1.3), it follows by Lemma A.61 that $\pi_k\Big(\mathbb{G}_S(\boldsymbol{\theta}_0)\Big)$ converges weakly to $\mathbb{G}_S(\boldsymbol{\theta}_0)$ as $k \to \infty$.

Now, $\nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_k}) \to \nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)$ as $k \to \infty$ follows immediately from Corollary A.16 under Assumption 3.5.3. The same holds true for $\nabla\Delta_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \to \nabla\Delta_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ as $k \to \infty$ for every $\boldsymbol{\theta} \in \text{lin}(\Theta)$. We thus obtain that $\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0) \to \boldsymbol{\beta}_\nabla^{\mathbb{S}_\Theta}(\boldsymbol{\theta}_0)$ and $\nabla\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \to \nabla\boldsymbol{\beta}_\nabla^{\mathbb{S}_\Theta}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \, \forall \boldsymbol{\theta} \in \text{lin}(\Theta)$ as $k \to \infty$. Under Assumption 3.5.8 we have that $\nabla\mu_k^\nabla \to \nabla\mu_\infty^\nabla$ as $k \to \infty$. Hence, it follows that,

$$\nabla Q_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \to \nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \, \forall \boldsymbol{\theta} \in \text{lin}(\Theta),$$

where $\nabla Q_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) := \nabla\mu_k^\nabla\Big(\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0) \, , \nabla\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})\Big)$. By continuity of the inverse operator and the continuous mapping theorem it thus follows that,

$$\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})\Big) \to \text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \boldsymbol{\theta})\Big) \, \forall \boldsymbol{\theta} \in \Theta.$$

Furthermore, by Lemma B.9 and Corollary B.11, $\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \cdot)\Big)$ is a bounded linear functional for every $k \in \mathbb{N}$. Hence, it follows that,

$$\text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_{\Theta_k}}}(\boldsymbol{\theta}_0, \cdot)\Big) \to \text{inv}\Big(\nabla Q_\infty^{\nabla_{\mathbb{S}_\Theta}}(\boldsymbol{\theta}_0, \cdot)\Big) \text{ as } k \to \infty \text{ uniformly on compact sets.}$$

The desired results now follows by the ECMT (Lemma A.54) and tightness, and by noting that,

$$\Psi_k := -\text{inv}\Big(\nabla\mu_k^\nabla\Big(\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0) \, , \nabla\boldsymbol{\beta}_\nabla^{\mathbb{S}_{\Theta_k}}(\boldsymbol{\theta}_0, \cdot)\Big)\Big)\Big(\pi_k\Big(\mathbb{G}_S(\boldsymbol{\theta}_0)\Big)\Big)$$

□

Chapter 4

Finite Sample Properties of SNPII Estimators

This chapter reports the results of a Monte Carlo exercise that provides a first description of the small sample properties of the SNPII estimator. Since SNPII estimation can easily become computationally very expensive, the examples covered here are mainly of a simple prototypical nature. Much work can still be done in studying the behavior of the SNPII estimator under alternative models, different sieves, richer sets of auxiliary statistics and with other criterion functions that might yield possibly different results.

Below, we analyze the small sample dehavior of the SNPII estimator in both a cross-sectional setting with independent identically distributed (iid) data and in a dynamic time-series econometric model.

In particular, Section 4.1 describes the SNPII estimator to be adopted in the simple cross-sectional regression setting. Section 4.2 proceeds to provide Monte Carlo evidence of the behavior of this estimator in the context of a regression with exponential conditional expectation function. This example provides simple albeit important insights into the workings and properties of SNPII estimation. In a dynamic setting, Section 4.3 describes the SNPII estimator to be used in the estimation of an econometric model derived from economic theory. Section 4.4 delivers the Monte Carlo results and reveals the importance of SNPII estimation even in restrictive settings with a small number of observations and 'small' sieves. Finally, Section 4.5 concludes.

4.1 Basic Formulation for Cross-Sectional Regression Models

Consider the use of SNPII estimators in the context of cross-sectional regression models. We assume to have at our disposal an iid sample $(y_1, x_1), ..., (y_T, x_T)$ of points $(y_t, x_t) \in \mathbb{R}^2$ drawn from the joint distribution of y and x. Our interest lies in estimating the conditional expectation function of y given x, denoted $\boldsymbol{\theta}_0(x) \equiv E(y|x)$. We assume that y and x are related according to,

$$y_t = \boldsymbol{\theta}_0(x_t) + \epsilon_t \tag{4.1}$$

where the error term $\epsilon_1, ..., \epsilon_T$ is (for simplicity) assumed to be iid with known distribution. Since $\boldsymbol{\theta}_0$ is unknown, the model postulated by the researcher takes the form,

$$\tilde{y}_t = \boldsymbol{\theta}(\tilde{x}_t) + \epsilon_t \tag{4.2}$$

with $\boldsymbol{\theta}$ allowed to be an element of the infinite dimensional parameter space Θ. For every T, the SNPII estimator $\hat{\boldsymbol{\theta}}_T$ is formulated so as to minimize a criterion function Q_T over the sieve $\Theta_T \subset \Theta$ with each sieve specified so as to satisfy $\Theta_T \subseteq \Theta_{T+1} \subseteq \Theta$ (see Chapters 2 and 3 for more details). In what follows the criterion function Q_T assumes a weighted quadratic form as in Gourieroux et al. (1993),

$$Q_{T,S}(\boldsymbol{\theta}) = \mu_T\Big(\hat{\boldsymbol{\beta}}_T\,,\, \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) = \sum_{i \in \mathbb{N}} w_{T,i}\Big(\hat{\boldsymbol{\beta}}^i_T - \tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta})\Big)^2. \tag{4.3}$$

The weights $w_{T,i}$ are chosen so as to be positive for the first k_T auxiliary statistics, and zero otherwise. The variable k_T is allowed to diverge to infinity with sample size at an appropriately chosen rate. For finite T and an appropriate choice of $w_{T,i}$, the criterion function thus boils down essentially to that of the parametric indirect inference estimator proposed in Gourieroux et al. (1993),

$$Q_T(\boldsymbol{\theta}) = \Big(\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big)\hat{W}_T\Big(\hat{\boldsymbol{\beta}}_T - \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big)' \tag{4.4}$$

where $\hat{W}_T$ might denote an estimator of an optimal weighting matrix W.[1] Recall from Chapters 2 and 3 the construction of $\hat{\boldsymbol{\beta}}_T$ as a vector of $\hat{\boldsymbol{\beta}}^i_T$'s and of $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ as a vector of averages $1/S \sum_{s=1}^{S} \tilde{\boldsymbol{\beta}}^i_{T,s}$ of corresponding auxiliary estimators $\tilde{\boldsymbol{\beta}}^i_{T,s}$. The auxiliary estimators $\hat{\boldsymbol{\beta}}^i_T$ are chosen to be least-squares estimators obtained by regressing y_t over $\mathcal{T}_i(x_t)$ for $i = 1, ..., k_T$, where $\mathcal{T}_i(x)$ denotes the i^{th} order Chebyshev

[1] Recall from Chapter 3 that in the context of SNPII estimation, the weights $w_{T,i}$ must satisfy some asymptotic properties. In particular, they must be chosen so as to guarantee convergence of $Q_T(\boldsymbol{\theta})$ to a well defined limit $\forall \boldsymbol{\theta}$.

polynomial of the first kind (Chapter 1). The same applies naturally to the $\tilde{\boldsymbol{\beta}}^i_{T,s}$'s making use of S streams of simulated data obtained from (4.2),

$$\hat{\boldsymbol{\beta}}^i_T = \arg\min_{\boldsymbol{\beta}_i \in \mathcal{B}_i} \sum_{t=1}^{T} \left(y_t - \boldsymbol{\beta}_i \mathcal{T}_i(x_t) \right)^2 \quad \text{and} \quad \tilde{\boldsymbol{\beta}}^i_{T,s} = \arg\min_{\boldsymbol{\beta}_i \in \mathcal{B}_i} \sum_{t=1}^{T} \left(\tilde{y}^s_t - \boldsymbol{\beta}_i \mathcal{T}_i(\tilde{x}^s_t) \right)^2. \tag{4.5}$$

This choice of auxiliary estimators is quite arbitrary. In essence, it is only important that the vectors $\hat{\boldsymbol{\beta}}_T$ and $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$ convey information about $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$ respectively. Depending on the nature of the problem, different choices of auxiliary estimators might be preferred.

In the present context, any alternative auxiliary estimators could be devised exploring nonlinearities and asymmetries in the dependence between y and x. A number of alternatives have been tested. These provided virtually identical results. As such, we stick to the choice in (4.5) which seems quite intuitive as a way of describing the nonlinear relationship between y and x.

Finally, a few words on the numerical/computational aspects of this study. Simulations were performed using the software package MATLAB. Given the heavy computational requirements of the SNPII estimator, the number of Monte Carlo replications in all results reported here is kept at the rather low (and computationally feasible) $N = 500$. For every replication, one set of artificial "observed data" was used to obtain an estimate of $\hat{\boldsymbol{\beta}}_T$ and $S = 10$ sets of simulated data were used to obtain an estimates of $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$. Clearly, the variance of the SNPII estimator could still be reduced by selecting a larger S. This choice seems to provide a good compromise between variance and computational requirements.

The sieves considered in this Monte Carlo exercise are of the linear type, i.e. they are a linear span of basis functions $\{\psi_1, ..., \psi_{k_T}\}$. Elements $\boldsymbol{\theta} \in \Theta_T$ are thus obtained as,

$$\boldsymbol{\theta}(x) = \sum_{i=1}^{k_T} \theta^i \psi_i(x)$$

where $\theta^1, ..., \theta^{k_T}$ are scalar parameters. Estimates $\hat{\boldsymbol{\theta}}_T$ of the conditional expectation function $\boldsymbol{\theta}_0$ are obtained through the estimation of the scalar parameters $\theta_1, ..., \theta^{k_T}$. In practice, these were obtained by minimizing the criterion function in (4.3) using a standard Newton-type algorithm. In every repetition, the initial parameter vector $(\theta^1, ..., \theta^{k_T})$ is just a vector of zeros. This corresponds naturally to an initial vector $\boldsymbol{\theta}$ in Θ that corresponds to zero function $\boldsymbol{\theta}(x) = 0 \, \forall x$. Alternative initial conditions seem to provide essentially identical results in this simple regression setting.

4.2 Monte Carlo Evidence from Simple Exponential Regression

For simplicity, consider the following trivial prototypical case. Let x_t take values uniformly on the interval $[3, 7]$, for every $t = 1, ..., T$. Suppose that $\boldsymbol{\theta}_0(x) \equiv E(y|x) = \exp(x)$ and hence that $y_t = \exp(x_t) + \epsilon_t$ holds for every $t = 1, ..., T$. Assume also for simplicity that $\epsilon_t \sim N(0, \sigma)$ with known $\sigma = 17$. Under such conditions, our interest lies in approximating $\boldsymbol{\theta}_0$ on the compact interval $[3, 7]$.[2] Given our particular choice of $\boldsymbol{\theta}_0$, we can now generate samples $(y_1, x_1), ..., (y_T, x_T)$ of "observed data" that specify a relation between y and x characterized by a conditional expectation taking the form of an exponential function.

Figure 4.1 shows a typical scatter plot for $T = 50$ obtained in our Monte Carlo exercise. Given the disturbance introduced by ϵ_t, it is not so trivial to identify with precision what the conditional expectation function of y given x might be. Based on visual inspection alone, one could be in fact tempted to suppose that y and x are simply linearly related. A fundamental step in the formulation of any estimation procedure is thus concerned with the choice of Θ. Typically, since $\boldsymbol{\theta}_0$ is unknown, there is *a priori* no reason to suppose that a 'small' Θ satisfies the typical axiom of correct specification $\boldsymbol{\theta}_0 \in \Theta$. Suppose for a moment that we postulate a linear relation between the covariates y and x; i.e. suppose that we are convinced that $E(y_t|x_t) = \theta^0 + \theta^1 x_t$, and thus restrict Θ to be a space of linear functions. In this case our Θ is spanned by the basis vectors $\{1, x\}$ and clearly $\boldsymbol{\theta}_0 \notin \Theta$.

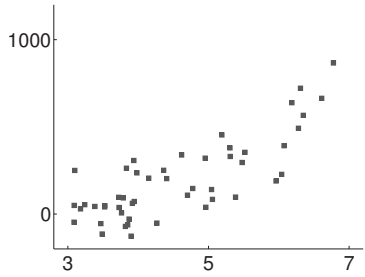

Figure 4.1: Typical scatter plot for $T = 50$ with $y_t = \boldsymbol{\theta}_0(x_t) + \epsilon_t$ and $\boldsymbol{\theta}_0(x) = \exp(x)$.

For any given $\boldsymbol{\theta} \in \Theta$, we can now generate samples $(\tilde{y}_1, \tilde{x}_1), ..., (\tilde{y}_T, \tilde{x}_T)$ of "simulated data" to be used in our indirect inference procedure. Estimates of the linear function of interest are obtained in the usual way by estimating the parameter vector $(\theta^0, \theta^1) \in \mathbb{R}^2$. Clearly, estimation of (θ^0, θ^1) does not require the aid of the indirect inference apparatus and we could proceed with classical direct estimation. Furthermore, the use of the auxiliary estimators described in (4.5) results in an indirect estimator that is characterized by a very simple and analytically tractable binding function, so no simulations are required. However, for the sake of argument and comparability with the results that follow, we make use of the simulation-based

[2] Examples dealing with unbounded intervals could also be considered.

indirect inference procedure described above. The small sample behavior of $\hat{\boldsymbol{\theta}}_T$ is shown in Figure 4.2 which plots $\boldsymbol{\theta}_0$ (in red) and the density of both $\hat{\boldsymbol{\theta}}_T$ and its deviation from $\boldsymbol{\theta}_0$ for a sample size $T = 50$.

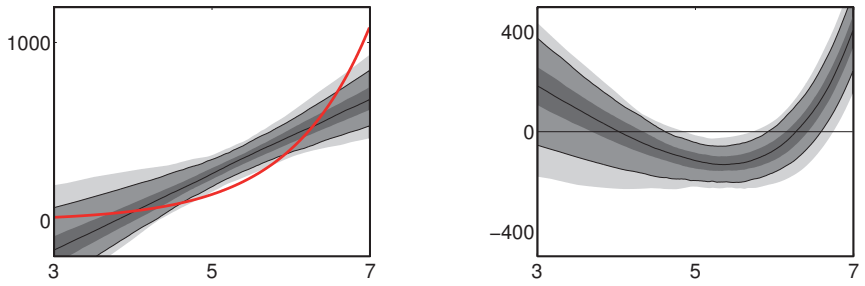

Figure 4.2: Density of $\hat{\boldsymbol{\theta}}_T$ (left) and $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0$ (right) for $T = 50$. The function in red is $\boldsymbol{\theta}_0$ (left) where $\boldsymbol{\theta}_0(x) = \exp(x)$. The parameter space Θ is the space of linear functions. The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75%.

Due to the incorrect specification of the regression model imposed by the restrictive Θ, our estimator will never converge to the appropriate limit $\boldsymbol{\theta}_0$. Figure 4.3 shows that as the sample size increases, a reduction in the variance of $\hat{\boldsymbol{\theta}}_T$ occurs. However, estimates of $\boldsymbol{\theta}_0 = E(y|x)$ are bound to produce unsatisfactory results.

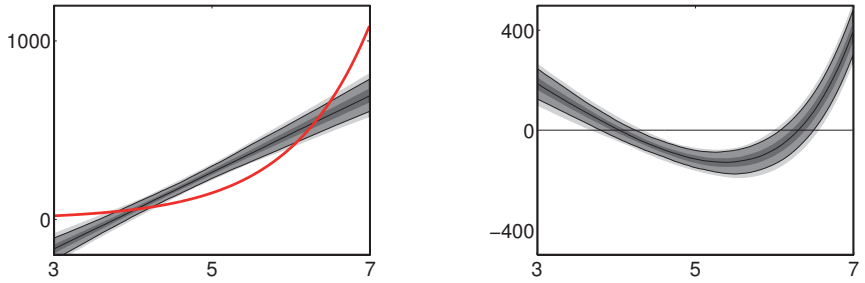

Figure 4.3: Density of $\hat{\boldsymbol{\theta}}_T$ (left) and $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0$ (right) for $T = 250$. The function in red is $\boldsymbol{\theta}_0$ (left) where $\boldsymbol{\theta}_0(x) = \exp(x)$. The parameter space Θ is the space of linear functions. The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75% of probability mass.

It is interesting to observe that even under the presence of misspecification, $\hat{\boldsymbol{\theta}}_T$ does not seem to behave erratically. This occurs because *(i)* the auxiliary estimators still converge in an appropriate fashion to a well defined singleton limit, and *(ii)* the binding function is injective.

The injective nature of the binding function is easy to ascertain in the present context. The convergence of the auxiliary estimators (under the regularity conditions

on the auxiliary parameter space introduced in Chapters 2 and 3) will be studied in Chapter 5.

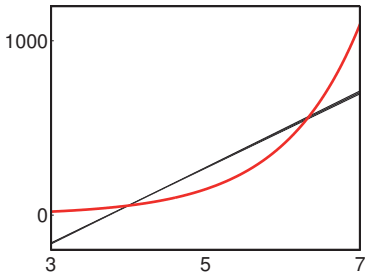

Figure 4.4: Density of $\hat{\boldsymbol{\theta}}_T$ for $T = 20000$ with $\boldsymbol{\theta}_0$ in red and Θ containing only linear functions. The probability mass is concentrated on a very small set of linear functions.

Indeed, the present model provides a very pragmatic example of how the results from approximation theory introduced in Chapter 5 are useful in determining the existence of a well separated minimizer of the least squares problem solved by the auxiliary statistics $\hat{\boldsymbol{\beta}}^i_T$ and $\tilde{\boldsymbol{\beta}}^i_{T,s}$. Figure 4.4 shows evidence of the convergence of $\hat{\boldsymbol{\theta}}_T$ to a well defined limit in the space of linear functions Θ. These considerations lead quite naturally to the conclusion that $\hat{\boldsymbol{\theta}}_T$ is converging to some point $\boldsymbol{\theta}^*_0$ in the space of linear functions Θ.

Remark 4.2.1. *As discussed in Section 1.2, $\boldsymbol{\theta}^*_0$ might have interesting properties and the interpretation of an* indirect pseudo-true parameter*. Indeed, $\boldsymbol{\theta}^*_0$ is at least (by construction) the minimizer of a divergence between the distribution of observed and simulated data (as measured by the auxiliary statistics and the criterion divergence μ_∞). If the auxiliary statistics are well chosen, $\boldsymbol{\theta}^*_0$ might thus be quite meaningful.*

Our concern here is however turned to the estimation of $\boldsymbol{\theta}_0$, not of some approximation $\boldsymbol{\theta}^*_0$. This is an impossibility in the present context and $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_\Theta$ does not vanish. Figure 4.5 provides evidence of this trivial fact.

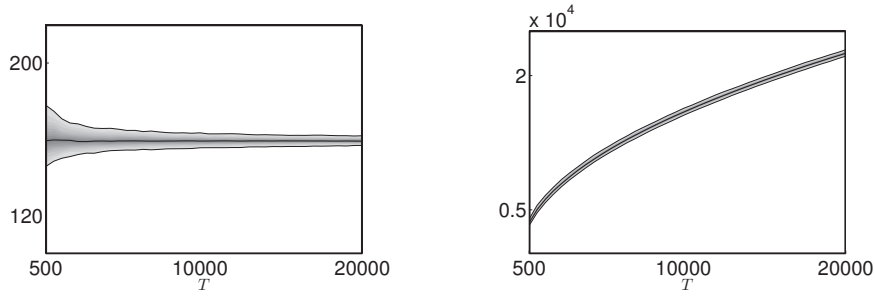

Figure 4.5: Density of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ (left) and $\sqrt{T}\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ (right) where $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2 = \left(\int |\hat{\boldsymbol{\theta}}_T(x) - \boldsymbol{\theta}_0(x)|^2 dx\right)^{1/2}$ is the L_2-norm, $\boldsymbol{\theta}_0(x) = \exp(x)$ and Θ contains only linear functions. Shaded area contains 95% of probability mass.

The inconsistency results reported until now are caused by the restrictive choice of Θ. Apparently, having Θ be spanned by the basis vectors $\{1, x\}$ does not provide us with a rich enough specification of the parameter space, and hence, $\boldsymbol{\theta}_0 \notin \Theta$. In

the spirit of parametric models one could proceed by adding a quadratic term to turn Θ into the richer space of real-valued quadratic functions, i.e. $\Theta = \text{lin}(\{1, x, x^2\})$. Since such a space contains that of linear functions, we are sure to have enlarged the possibilities of finding $\boldsymbol{\theta}_0$. Unfortunately, we know already that this will not provide a solution to our problem with $\boldsymbol{\theta}_0(x) = \exp(x)$. Indeed, the fundamental problem faced here by the use of the parametric indirect inference procedure is really that misspecification remains, no matter how many power monomials we add to our basis vector. Figures 4.6 and 4.7 reveal the results obtained under for the indirect inference estimator that postulates a quadratic regression model.

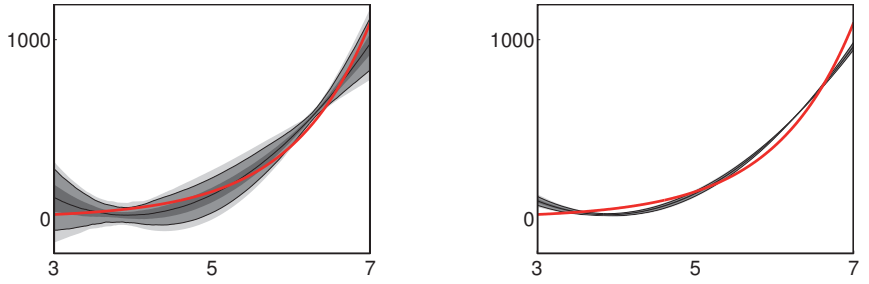

Figure 4.6: Density of $\hat{\boldsymbol{\theta}}_T$ for $T = 500$ (left) and $\hat{\boldsymbol{\theta}}_T$ for $T = 20000$ (right). The function in red is $\boldsymbol{\theta}_0$ (left) where $\boldsymbol{\theta}_0(x) = \exp(x)$. The parameter space Θ is the space of quadratic functions. The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75%.

Figure 4.6 suggests that $\hat{\boldsymbol{\theta}}_T$ converges once more to a well defined limit as $T \to \infty$. This time, a quadratic function $\boldsymbol{\theta}_0^* \in \Theta$.[3] Figure 4.7 suggest furthermore the existence of a substantial gain in terms of accuracy compared to the linear model as measured by the L_2-norm. There is however no hope of obtaining a consistent estimate of $\boldsymbol{\theta}_0$ and $\sqrt{T}\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ diverges.

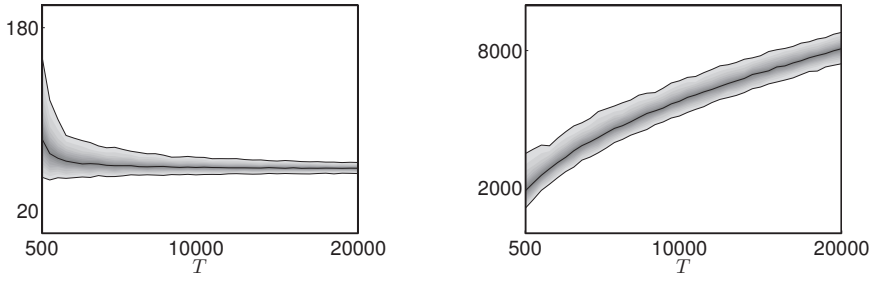

Figure 4.7: Density of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ (left) and $\sqrt{T}\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ (right) where $\|\boldsymbol{\theta}\|_2 = \left(\int |\boldsymbol{\theta}(x)|^2 dx\right)^{1/2}$ is the L_2-norm, $\boldsymbol{\theta}_0(x) = \exp(x)$ and Θ contains only quadratic functions. Shaded area contains 95% of probability mass.

[3]Once again this can be shown by appealing to the results in Chapter 5.

Let us finally turn to a more promising approach of obtaining consistent estimates of $\boldsymbol{\theta}_0$. For comparison with the previous results, we retain the power monomials as a source of basis vectors. This time however, we adopt the SNPII estimation procedure laid down in Chapters 2 and 3 and let our estimator $\hat{\boldsymbol{\theta}}_T$ take values in sets Θ_T obtained as $\Theta_T = \text{lin}(\{1, x, x^2, ..., x^{k_T}\})$ with $k_T \to \infty$ as $T \to \infty$.

Figure 4.8 plots fast growing truncation order k_T that satisfies $k_T = O(T^{1/3})$. That particular choice implies a linear regression model for $T \leq 90$, a quadratic regression for $90 < T \leq 166$, a cubic regression for $166 < T \leq 275$, and so on.

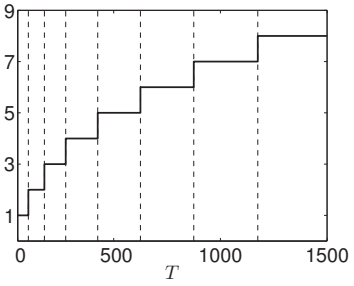

Figure 4.8: Plot of $k_T \approx T^{1/3} - 3$.

The consistency theorems in Chapters 2 and 3 postulate that, under appropriate regularity conditions, $\hat{\boldsymbol{\theta}}_T$ will be consistent to any $\boldsymbol{\theta}_0$ lying in a space Θ whose elements are arbitrarily well approximated by a sequence in $\{\Theta_T\}_{T\in\mathbb{N}}$. Weierstrass's Theorem (Lemma A.92) thus suggests that consistency might be obtained for every continuous $\boldsymbol{\theta}_0$. Further restrictions must however be imposed if we wish $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|$ to be $O(T^{-1/2})$ and for $\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)$ to converge to the limit distribution obtained in Chapters 2 and 3. Indeed, recall that for $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|$ to vanish at an appropriate speed, a minimum expansion rate on the sieves Θ_T must be imposed.

In the present context, we make use of results on the speed with which truncated power series of increasing order approximate certain classes of functions. In general, such results will relate the speed at which the truncation order k_T diverges with an upper bound on the rate at which $\inf_{\boldsymbol{\theta}\in\Theta_k} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\Theta$ vanishes to zero (under some norm $\|\cdot\|_\Theta$).

Remark 4.2.2. *If Θ is taken to be the Hölder space $\mathcal{H}^p(\mathcal{X})$ of p-smooth functions on a compact subset $\mathcal{X}$ of $\mathbb{R}^d$, then, for any $\boldsymbol{\theta}_0 \in \Theta$, approximation from a sieve of k_T^{th}-order polynomials $\Theta_{k_T} = \text{lin}\{1, x, ..., x^{k_T}\}$ satisfies $\inf_{\boldsymbol{\theta}\in\Theta_k} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty = O(k_T^{-p/d})$.*[4] *See Powell (1981), Judd (1998) and Chen (2007) for various similar results.*

In the context of our exponential regression, a k_T satisfying $k_T = O(T^{1/5})$ is thus capable of approximating any $\boldsymbol{\theta}_0 \in \mathcal{H}^p([3, 7])$, for any $p > 5/2$, in any L_{p^*}-norm at the desired rate since $\inf_{\boldsymbol{\theta}\in\Theta_k} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty = o(T^{-1/2})$ and naturally $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{p^*} \leq$

[4] $\|\cdot\|_\infty$ denotes the sup-norm. Hence, the results apply naturally to other L_p-norms. Furthermore, note that the Hölder space of p-smooth functions is composed of the m-times differentiable functions with m^{th} derivative satisfying a γ-Hölder continuity condition and $p = m + \gamma$; see e.g. Chen (2007).

$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty$ for any $p^* \in \mathbb{N}$. With $k_T = O(T^{1/3})$ approximation at appropriate rates is extended to the larger parameter spaces $\Theta \equiv \mathcal{H}^p([3,7])$ for any $p > 3/2$.

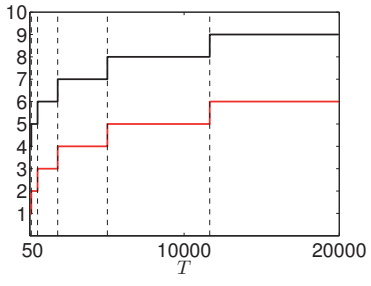

Figure 4.9: Plot of $k_T \approx T^{1/5}+2$ (black line) and $k_T \approx T^{1/5} - 1$ (red line).

Figure 4.9 presents two alternative choices of k_T, both satisfying $k_T = O(T^{1/5})$. Asymptotically, these are equivalent. In particular, under the appropriate regularity conditions, both ensure the $\sqrt{T}$-convergence and asymptotic Gaussianity of $\hat{\boldsymbol{\theta}}_T$. In applications however, the choice matters and might lead to quite different results. With a relatively small sample of $T = 90$ observations, while one sequence k_T (red line) implies the use of a simple linear regression, the other sequence k_T (in black) implies the estimation of a quartic regression. Clearly, when compared to the former, the latter regression will exhibit both strengths and weaknesses. On the one hand, the estimation of a quartic regression requires the estimation of a larger number of parameters (associated to 1, x, x^2, x^3 and x^4). This implies the natural increase of estimation uncertainty and associated computation of larger confidence intervals. On the other hand, the increased flexibility in the regression, associated with the adoption of a larger sieve Θ_T, is likely to reduce the finite sample bias associated with the sieve's restriction and might result in the derivation of an *approximate asymptotic distribution* which is closer to the exact asymptotic distribution of the SNPII estimator (in Theorem 3.5.1 of Chapter 3).

Figure 4.10 presents an alternative case, with a choice of k_T satisfying $k_T = O(T^{1/3})$ (black line) and another where $k_T = O(T^{1/5})$ (red line). Both choices imply a linear regression for $T = 50$. As T grows, the sieves grow at different rates. This time, these choices carry important asymptotic implications. In particular, the results of Chapter 2 and 3 suggest that for $\boldsymbol{\theta}_0 \in \mathcal{H}^2([3,7])$, correct convergence rates will only be obtained for the faster growing k_T. As mentioned above, the slowest k_T allows us to conduct statistical inference based on asymptotic arguments only on a smaller space e.g. $\mathcal{H}^3([3,7])$.

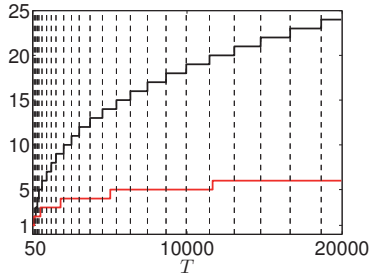

Figure 4.10: Plot of fast increasing $k_T \approx T^{1/3} - 3$ slowly increasing $k_T \approx T^{1/5} - 1$.

Let us now analyze more carefully the implications of such choices. Figure 4.11

shows evidence of how the SNPII estimator behaves for increasing sample sizes for the fast growing truncation order $k_T = T^{1/3} - 3$ plotted in Figure 4.8.

These graphs document well the effects of sieve restrictions. For $T = 50$, the SNPII estimator takes values in a sieve of linear functions. As T grows, $\hat{\boldsymbol{\theta}}_T$ is then allowed to become more flexible and its density becomes tighter around $\boldsymbol{\theta}_0$.

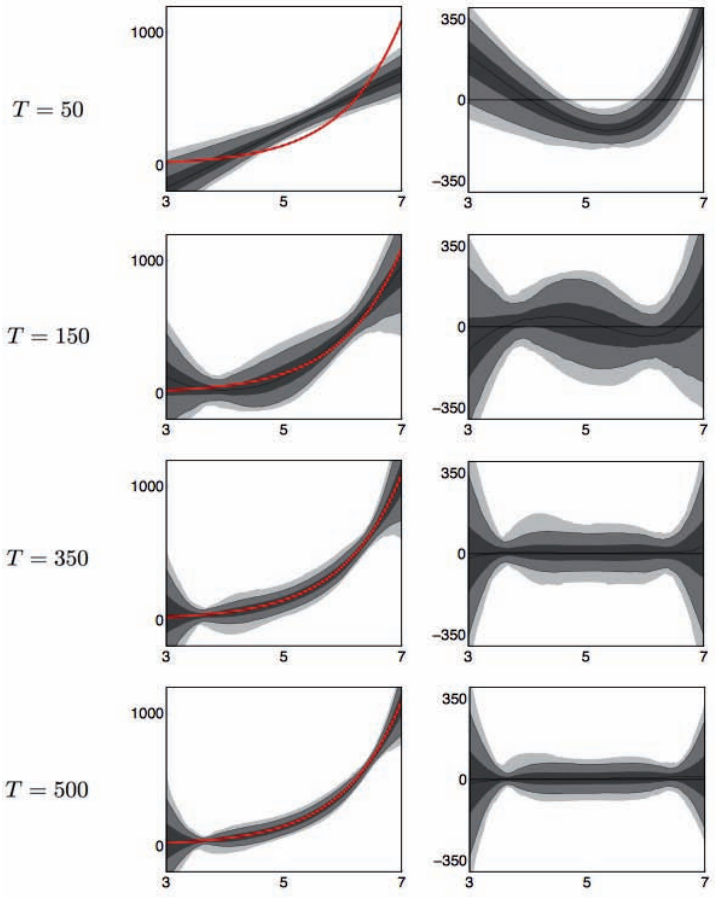

Figure 4.11: Density of the SNPII estimator $\hat{\boldsymbol{\theta}}_T$ (left column) and $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0$ (right column) for increasing sample size ranging from $T = 50$ to $T = 500$. The function in red is $\boldsymbol{\theta}_0$ (left column) where $\boldsymbol{\theta}_0(x) = \exp(x)$. The sieves are obtained as $\Theta_T = \text{lin}\{1, x, x^2, ..., x^{k_T}\}$ with slowly growing $k_T \approx T^{1/3} - 3$ (black line in Figures 4.8 and 4.10). The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75% of probability mass.

As already pointed out, the advantage of the fast growing truncation order k_T is one of generality. In this particular example however, where $\boldsymbol{\theta}_0(x) = \exp(x)$, it

is clear that k_T can be allowed to grow at a much slower rate. In particular, since we know here that $\boldsymbol{\theta}_0$ is analytic, we feel comfortable in adopting an alternative k_T, even if it comes at the cost of some generality. Figure 4.12 reveals Monte Carlo results for the SNPII estimator under the slow growing $k_T \approx O(T^{1/5})$ plotted in Figure 4.9 (red line).

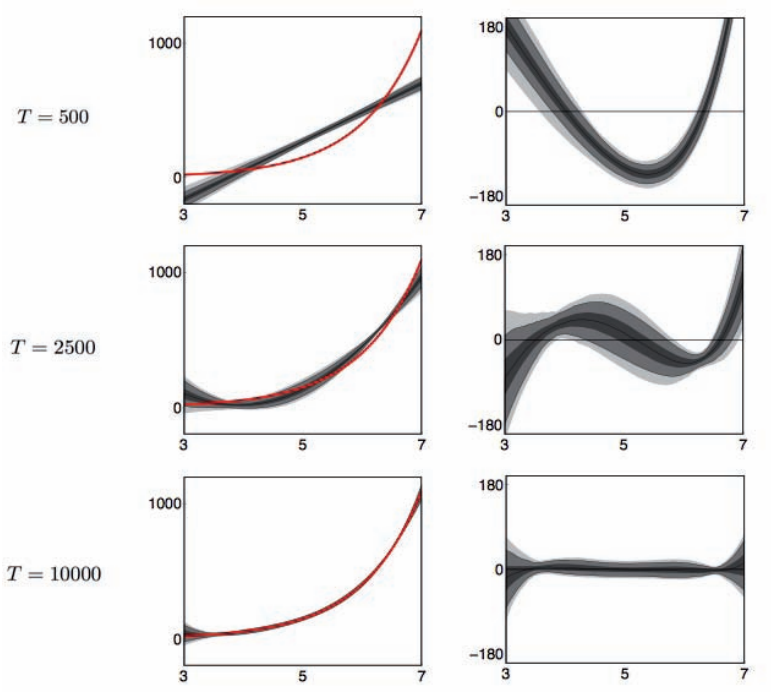

Figure 4.12: Density of $\hat{\boldsymbol{\theta}}_T$ (left column) and $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0$ (right column) for increasing sample size ranging from $T = 500$ to $T = 10000$. The function in red is $\boldsymbol{\theta}_0$ (left column) where $\boldsymbol{\theta}_0(x) = \exp(x)$. The sieves are obtained as $\Theta_T = \text{lin}\{1, x, x^2, ..., x^{k_T}\}$ with slowly growing $k_T \approx T^{1/5} - 1$ (red line in Figures 4.9 and 4.10). The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75%.

Again, the restrictions imposed by the sieve structure imposed on $\hat{\boldsymbol{\theta}}_T$ are strikingly clear. Note also that the chosen sample sizes in Figure 4.12 are now quite different from those adopted in Figure 4.11. Indeed, for the chosen $k_T \approx T^{1/5} - 1$, if we were to plot the density of $\hat{\boldsymbol{\theta}}_T$ for sample sizes of 50, 150 and 350 observations, we would simply observe the results of a linear regression model estimation comparable to those already reported in Figures 4.2 and 4.3. An interesting comparison of Figures 4.11 and 4.12 can nonetheless be made. Let us focus for a moment on the sample size $T = 500$. Visual inspection of both figures reveals the small sample trade-off between adopting more or less restrictive sieves. Notice in particular that

while the "larger" sieve (spanned by power monomials of up to fifth order) in Figure 4.11 allows $\hat{\boldsymbol{\theta}}_T$ to provide a seemingly better description of the "shape" of $\boldsymbol{\theta}_0$, the smaller sieve (of linear functions) in Figure 4.11 produces a $\hat{\boldsymbol{\theta}}_T$ with considerably lower variance.

Regardless of the divergence rate of k_T, the most interesting aspect of Figure 4.12 is that it testifies once again that the distribution of $\hat{\boldsymbol{\theta}}_T$ becomes increasingly tight around $\boldsymbol{\theta}_0$, as the sample size increases. In fact, it testifies that for a large $T = 10000$ the density of $\hat{\boldsymbol{\theta}}_T$ is already concentrated on a very small set of functions around $\boldsymbol{\theta}_0$. Recall that for $k_T = O(T^{1/5})$, the SNPII estimator is capable of approximating any function in $\mathcal{H}^p([3,7])$, for any $p > 5/2$, at appropriate rates. Since in our case, $\boldsymbol{\theta}_0(x) = \exp(x)$ we know for sure that $\boldsymbol{\theta}_0 \in \mathcal{H}^p([3,7])$ for any $p \in \mathbb{N}$. Hence, by the results of Chapters 2 and 3 we expect $\hat{\boldsymbol{\theta}}_T$ to be $\sqrt{T}$-consistent estimator, and thus to observe $\sqrt{T}\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2 = O(1)$. This can be seen quite clearly in Figure 4.13.

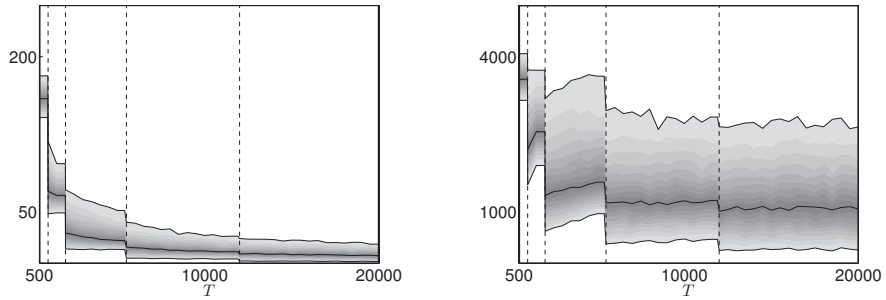

Figure 4.13: Density of $\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ (left) and $\sqrt{T}\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\|_2$ (right) where $\|\boldsymbol{\theta}\|_2 = \left(\int |\boldsymbol{\theta}(x)|^2 dx\right)^{1/2}$ is the L_2-norm, $\boldsymbol{\theta}_0(x) = \exp(x)$ and Θ contains only quadratic functions. Shaded area contains 95% of probability mass. Dashed vertical lines indicate sample sizes at which k_T increases.

This simple regression example was quite instructive as it allowed us to obtain insightful Monte Carlo results while keeping computational requirements at an acceptable level. In the following section we finally turn to dynamic models. On the one hand, the results covered below will be more meaningful since they deal with a model derived from economic theory. On the other hand, the Monte Carlo exercise will deliver more limited results as the computational requirements are heavier.

4.3 Basic Formulation for Dynamic Models

In this section we analyze briefly the use of SNPII estimators in the context of dynamic models. In particular, we suppose that observed data consists of observations from a vector time-series process.

In the absence of economic-theoretic restrictions, autoregressive models of the

type $\mathbf{x}_t = \boldsymbol{\theta}_0(\mathbf{x}_{t-1}, \boldsymbol{\epsilon}_t)$ similar to that considered in the very beginning of this thesis (in Section 1) are often interesting. In such cases, SNPII estimation should proceed essentially as in the previous section. Indeed, except for a number of details concerning the appropriate choice of auxiliary estimators and sieves, everything else applies essentially in the same way. Most importantly, as suggested in Section 1.4, choices of sieves appropriate for dynamic models (e.g. artificial neural networks) will allow $\boldsymbol{\theta}_0$ to be consistently estimated even when it lies in very general spaces (see also Granger and Terasvirta (1993)).

Remark 4.3.1. *In nonlinear dynamic models, auxiliary estimators must be chosen so as to describe appropriately the dynamic features of the data. Furthermore, as pointed out in Section 3.7, sieves should be selected so as to ensure that certain dynamic properties of interest such as stability and fading memory hold.*[5]

Below, we shall focus on making use of SNPII estimators in the context of theory-driven models. Our aim is to analyze the behavior of SNPII estimators when models are formulated *"in conjunction with appropriate theories"* as suggested by Granger and Terasvirta (1993). For concreteness, let us turn back to the basic RBC model considered in Section 1.6 of Chapter 1. This simple model of an isolated farm abstracts from the complications introduced by the larger and more complex models. Nonetheless, it might provide some important insight into the behavior of SNPII estimators in the context of dynamic theory-driven models in general.

Recall that in the context of Section 1.6, economic theory derives the dynamic behavior of economic variables from the following optimization problem,

$$\max_{\{c_t\}_{t=1}^{\infty}} E_t\left[\sum_{s=t}^{\infty} \beta^{s-t} u(c_s)\right] \quad , \quad \text{s.t.} \quad k_{t+1} = f(k_t, z_t) - c_t \, , \quad z_t = g(z_{t-1}) + \epsilon_t. \tag{4.6}$$

Certain features of this optimization problem might be better described by economic theory than others, in which case economists will be more confident of some theoretic restrictions than others. For example, the general description of capital accumulation of the 'isolated farmer' as the process through which a quantity of cereals is consumed while the remaining stock is used for obtaining new crops in the next season, might be consensual and accurate. Likewise, the general specification of the TFP has having some form of time dependence as described by a nonlinear autoregressive process might be deemed reasonable. However, there is in general, great uncertainty as to which exact form the functions, u, f and g might take. This difficulty is generally accepted in economics.

[5] Recall that catalogues of conditions for the formulation of nonlinear dynamic models with appropriate fading memory including near epoch dependence and L_0-approximability conditions can be found in Gallant and White (1988b) and Pötscher and Prucha (1997). Also, Trapletti et al. (1998) provide geometric ergodicity and stationarity conditions directly for artificial neural network sieves.

> *"One of the main differences between econometrics and the application of statistical methods in the physical sciences is that the functional forms in the structural equations of an econometric model are seldom given by the theory"* in Bergstrom (1985).

Economic theory is often capable of providing very general conditions under which utility functions or production functions are continuous, monotone, concave, etc. (see e.g. Debreu (1959)). These very general results are however far from the restrictions that are typically imposed in empirical work. Unfortunately, in the face of such difficulties, it is common for researchers to proceed by parametrizing the unknown functions u, f and g according to very simplistic (and thus restrictive) forms. Common examples consist of CRRA utility functions $u(c_t) = c_t^{1-\theta_u}/(1-\theta_u)$, AK production functions $f(k_t, z_t) = \exp(z_t)Ak_t^{\theta_f}$ and linear TFP equations $g(z_{t-1}) = \theta_g z_{t-1}$. The choice of such functions is typically justified in an informal way by the desire to retain algebraic convenience and analytical simplicity. For example, Kydland and Prescott (1982) give the following justification for the form of a production function featured in their RBC model.

> *"The production function is assumed to have the form* $f(\lambda, k, n, y) = \lambda n^{\theta}[(1-\sigma)k^{-v} + \sigma^{-v}]^{-(1-\theta)/v}$ *where* $0 < \theta < 1$, $0 < \sigma < 1$, *and* $0 < v < \infty$. *This form was selected because, among other things, it results in a share* θ *for labor in the steady state."* in Kydland and Prescott (1982).

It is thus not surprising to find proponents of such theory-driven models as Lucas (1985) concluding that *"Of course, the model is not 'true' "*. Under these typical restrictive assumptions, the system of dynamic first-order conditions derived from the optimization problem above, takes the form,

$$\begin{aligned} c_t^{-\theta_u} &= \beta E_t\left[c_{t+1}^{-\theta_u}\theta_f \exp(z_{t+1})Ak_{t+1}^{\theta_f - 1}\right] \\ k_{t+1} &= \exp(z_t)Ak_t^{\theta_f} - c_t \\ z_t &= \theta_g z_{t-1} + \epsilon_t\ ,\ \ \epsilon_t \sim N(0, \sigma^2) \end{aligned}$$

Finally, under additional conditions involving rational expectations of agents, the system of first-order conditions above is then 'solved' (and typically linearized in the process) and turned into a system of dynamic autoregressive equations determining the behavior of the variables of interest. This dynamic system is then the focus of econometric analysis. However, in the likely event of model misspecification, econometric analysis can be misleading.

Remark 4.3.2. *It is precisely the informal justification of functional form found in the quote above from Kydland and Prescott (1982) that SNPII theory helps to avoid.*

In essence, SNPII theory exploits the vast body of mathematical results on approximation theory to assist economist and econometricians on the design of functional forms that have greater chances of being coherent with general theory, empirical observation and correct specification axioms.

An example of how SNPII theory can be used in conjunction with economic theory consists precisely of letting Approximation Theory guide the process of choosing functional forms for u, f and g. In particular, sieves can be chosen *in conjunction with theory* so as to obtain an SNPII estimator that is indeed capable of consistently estimating any element within a class of functions suggested by theory, e.g. functions that are continuous, increasing, monotone, concave, and others.

A particular example consisting of polynomial approximations to these functions leads to a an optimization problem given by (4.6) where,[6]

$$u(c_t) \approx \sum_{i=0}^{k_T^u} \theta_{u,i}(c_t - c_{ss})^i \ , \qquad g(z_{t-1}) \approx \sum_{i=0}^{k_T^g} \theta_{g,i}(z_{t-1} - z_{ss})^i \ ,$$

$$f(k_t, z_t) \approx \sum_{i=0}^{k_T^f} \sum_{j=0}^{k_T^f} \theta_{f,i,j}(k_t - k_{ss})^i (z_t - z_{ss})^j \ .$$

In the spirit of SNPII estimation, important generality might be gained by letting the truncation orders k_T^u, k_T^g and k_T^f diverge to infinity at appropriate rates. As usual, a system of first-order conditions can once again be derived,

$$\sum_{i=1}^{k_T^u} i\theta_{u,i} c_{p,t}^{i-1} = \beta E_t \Big[\sum_{i=1}^{k_T^f} \sum_{j=0}^{k_T^f} i\theta_{f,i,j}(k_{t+1} - k_{ss})^{i-1}(z_{t+1} - z_{ss})^j \sum_{i=1}^{k_T^u} i\theta_{u,i} c_{p,t}^{i-1} \Big]$$

$$k_{t+1} = \sum_{i=0}^{k_T^f} \sum_{j=0}^{k_T^f} \theta_{f,i,j}(k_t - k_{ss})^i (z_t - z_{ss})^j - c_t$$

$$z_t = \sum_{i=0}^{k_T^g} \theta_{g,i} z_{t-1} + \epsilon_t \ , \ \epsilon_t \sim N(0, \sigma^2).$$

Finally, in the context of rational expectation models, SNPII estimation can proceed by applying appropriate nonlinear solution methods that approximate the consumption *policy function* with any desired level of accuracy. In particular, *perturbation methods* (see e.g. Judd Judd (1998)) might be especially well suited in this context, since they also approximate the policy function by a truncated power series. Various other methods, including spectral projection, finite element or spline methods, might be preferable depending on the choice of sieves. An issue that seems

[6]Here, c_{ss}, k_{ss} and z_{ss} denote steady-state quantities.

to have been ignored in this literature concerns however the fact that (just like polynomial sieves) perturbation solution methods do not confer the dynamic model with appropriate stability properties. The same applies to several other solution methods if appropriate conditions are not imposed.

Some limited simulation based experiences suggest nonetheless that the SNPII estimator works very well with Chebyshev spectral projection and perturbation solution methods. The computational requirements prevent us however from developing a fully fledged Monte Carlo exercise. In Section 4.4 we shall thus avoid the 'solution' part of the modeling process.

In what follows we keep working with a weighted quadratic criterion function Q_T as in Gourieroux et al. (1993),

$$Q_{T,S}(\boldsymbol{\theta}) = \mu_T\Big(\hat{\boldsymbol{\beta}}_T\ ,\ \tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})\Big) = \sum_{i\in\mathbb{N}} w_{T,i}\Big(\hat{\boldsymbol{\beta}}^i_T - \tilde{\boldsymbol{\beta}}^i_{T,S}(\boldsymbol{\theta})\Big)^2. \tag{4.7}$$

This time however, auxiliary estimators are chosen to be least-squares estimators obtained by regressing y_t over $\mathcal{T}_k(y_{t-1})$ for $k = 1, ..., k_T$, where $\mathcal{T}_k(y_{t-1})$ denotes the k-th order Chebyshev polynomial transformation of the lag y_{t-1}. Multiple lags are also considered. This choice seems to offer enough "information" about the nonlinear autoregressive structure of the data. As before, alternative auxiliary statistics exploring nonlinearities and asymmetries in the dependence between y and its lags seem to provide virtually identical results.

Simulations were once again performed using the software package MATLAB with a number of Monte Carlo replications of $N = 500$. For every replication, one set of artificial "observed data" was used to obtain $\hat{\boldsymbol{\beta}}_T$ and $S = 20$ sets of simulated data were used to obtain $\tilde{\boldsymbol{\beta}}_{T,S}(\boldsymbol{\theta})$. Finally, actual SNPII estimates $\hat{\boldsymbol{\theta}}_T$, were again obtained by minimizing the criterion function using a standard Newton-type algorithm. Alternative initial conditions seem to provide essentially identical results.

4.4 Monte Carlo Evidence from Simple Econometric Model

To obtain Monte Carlo results with little computational effort, let us further simplify the RBC model discussed above by considering only the dynamic equations that describe capital accumulation and TPF fluctuations over time.[7] In particular,

[7] This saves considerable time since solution methods for rational expectation models are avoided. Some limited experiences suggest that the SNPII estimator with Chebyshev spectral projection solutions, or high-order perturbation solutions (see e.g. Judd (1992, 1998)) work very well.

let us generate iid normal consumption sequences $\{c_t\}$ from $N(c_{ss}, \sigma_c)$ and obtain sequences of capital stock and TFP shocks according to,

$$\begin{aligned} k_{t+1} &= f(k_t, z_t) - c_t \\ z_t &= g(z_{t-1}) + \epsilon_t \quad , \qquad \epsilon_t \sim N(0, \sigma_\epsilon). \end{aligned} \tag{4.8}$$

The 'true' functions f and g are assumed to take quite common forms,

$$f(k_t, z_t) = (1 + z_t) A \log(k_t) \qquad \text{and} \qquad g(z_{t-1}) = \theta_g z_{t-1} \quad ,$$

so that the production function is linear in TFP shocks and concave in capital, and TFP dynamics are of the linear autoregressive type.[8]

The benefits of the econometric technique are better revealed by choosing a setting that is especially challenging for SNPII estimation. Below, we devise a Monte Carlo study that imposes such a setting. First, the sample size is restricted to a rather small $T = 200$. This is certainly problematic for SNPII estimators that essentially rely on increasing sample sizes and high dimensional parameter spaces to approximate complex DGPs. Second, in the spirit of most economic-theoretic research, we restrict the sieves to retain the extremely narrow simplicity of second-order polynomials. In particular, we keep working with sieves generated as $\Theta_T = \text{lin}\{1, x, ..., x^{k_T}\}$ yet impose that $k_T = 2$ for $T = 200$. Third, we compare the SNPII estimator to a standard parametric indirect inference estimator, supposing that the researcher has actually devised an impressively correct parametric description of the data generating process. In other words, we assume that economic theory has actually led the researcher to postulate a dynamic model almost identical to the actual DGP.

From now on, we suppose that the general form of the capital accumulation and the linear autoregressive structure of the TFP shocks has been accurately derived from theory (i.e. that the researcher works with the dynamic equations in (4.8)). Furthermore, we assume that the distribution of the shocks ϵ_t is also known. Finally, we also suppose that economic theoretic considerations have lead the researcher to accurately describe the production function has being smooth, monotonic and concave in capital and increasing in TFP. The only, deviation from the DGP concerns the exact form of f which the researcher postulates to be (the equally common) $\exp(z_t) A k_t^{\theta_f}$, $\theta_f \in (0, 1)$.[9] This production function is very similar to the 'true' one in the capital dimension, but convex instead of linear in the TFP dimension.

[8] The variance σ_ϵ of ϵ_t is selected small enough to ensure that $(1 + z_t) < 0$ occurs with negligible probability. This ensures in practice that negative production $(1 + z_t) A \log(k_t)$ does not occur in simulations. Likewise A is selected to have the steady-state of capital at $k_{ss} = 100$ so that $\log(k_t) < 0$ does not occur in practice. Finally, $\theta_g = 0.8$ yields temporal dependence to TFP shocks.

[9] The constant A is used essentially to determine the steady-state.

Remark 4.4.1. *Even under these rather favorable conditions for parametric estimation, the SNPII estimator can be shown to be of important value. It is important to note that the true production function* $f(k_t, z_t) = (1 + z_t)A\log(k_t)$ *is not an element of the space* $\mathcal{F}(k_t, z_t) = \{\exp(z_t)Ak_t^{\theta_f}\,,\ \theta_f \in (0, 1)\}$ *nor of the sieve of quadratic functions. So that, loosely speaking, both estimators below suffer from 'incorrect specification' at* $T = 200$.

Let us now analyze the Monte Carlo results. We focus on the estimates of the production function f and its derivatives.

Figure 4.14 below, plots the densities of both the classical indirect inference estimator (left) and the SNPII estimator (right) of the production function f along the capital dimension $k_t \in [80, 120]$ (at fixed steady-state TFP level of $z_{ss} = 0$).[10] Despite, 'small amounts of misspecification' and the 'almost linear shape' of $f(\cdot, z_{ss})$ on $[80, 120]$, the plots reveal that the variance of the SNPII estimator is considerably smaller than that of the standard II estimator. At the very minimum, this suggests the importance of adopting 'flexible' functional forms. More generally, it reveals the importance of SNPII estimation even under the restrictions imposed by small sample sizes and relatively simple sieves.

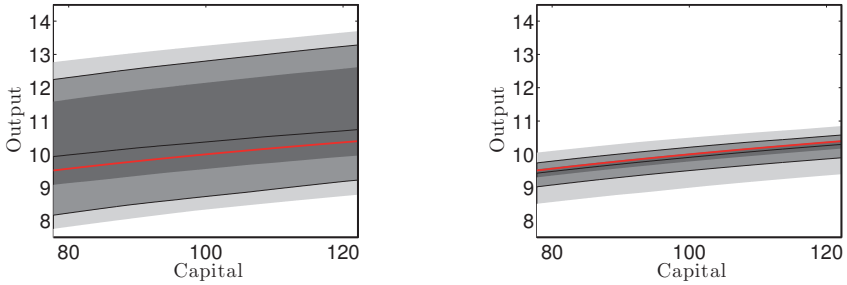

Figure 4.14: Densities of standard II estimator (left) and SNPII estimator (right) of f *along* $k \in [80, 120]$ *for fixed* $z_{ss} = 0$ *and* $T = 200$. *The function in red is the true production function* $f(k_t, z_{ss}) = (1 + z_{ss})A\log(k_t)$. *The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75%.*

The same analysis can be carried out along the TFP dimension. Figure 4.15 plots the densities of both the parametric II estimator (left) and the SNPII estimator (right) of the production function f along the TFP dimension $z_t \in [-1, 1]$ (at fixed steady-state capital level of $k_{ss} = 0$).[11] Here, as expected, the results are even more striking. This is due to the imposed linearity of f along the TFP dimension in the misspecified parametric model. Again, the Monte Carlo evidence suggests

[10] The choice of range values for capital stock of $[80, 120]$ is justified by the fact that such bounds are sufficiently large to contain virtually all paths of simulated capital stock from the DGP model.

[11] The interval for TFP $[-1, 1]$ contains virtually all paths of simulated TFP from the DGP model.

the importance of 'flexible' or 'exploratory' econometric techniques. In essence, the severe theoretical restrictions on the production function seem to result in possibly misleading statistical inference.

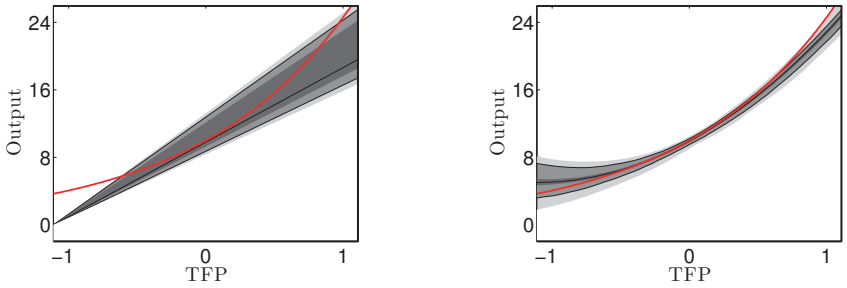

Figure 4.15: Densities of standard II estimator (left) and SNPII estimator (right) of f along $z \in [-1, 1]$ for fixed $k_{ss} = 100$ and $T = 200$. The function in red is the true production function $f(k_{ss}, z_t) = (1 + z_t)A\log(k_{ss})$. The light gray region contains 99% of the probability mass. The gray region contains 95%. The dark gray region contains 75%.

Finally, Figure 4.16 below, shows densities of estimated derivatives of the production function obtained using the standard parametric estimator (in dark grey) and those of the SNPII estimator (light grey).[12] True parameter values are marked by a red vertical line. Once again, despite the 'small amounts of misspecification' the evidence reveals that the standard parametric indirect inference estimator suffers from severe bias. In contrast, the SNPII estimator performs considerably better even under very restrictive sieves. The variance of the SNPII estimator is also considerably smaller. In fact, its distribution is considerably tighter around the true parameter values. Evidence of the existence of multiple 'pseudo-true' parameters supported by the lack of unimodality in the densities is also quite striking. Chapter 5 will analyze in more detail the conditions under which one might expect to have uniqueness of 'pseudo-true' parameters in misspecified models.

4.5 Final Remarks

This chapter provided a first account of the finite-sample behavior of SNPII estimators. The Monte Carlo exercises conducted here suggest the advantages of adopting flexible econometric techniques like SNPII estimation. On increasing sample sizes, the SNPII estimator seems to behave as expected and to converge at the appropriate rates to its limit. In the context of theory-driven models, the SNPII estimator seems also to have important advantages over other parametric estimators.

[12]Recall that the theory in Chapter 3 covered also the estimation of functionals such as derivatives.

Further research on the small sample properties of SNPII estimators should include a detailed analysis of alternative criterion functions, auxiliary estimators and sieves. Most importantly, in the context of dynamic models, a serious analysis of alternative sieves ensuring the appropriate stability and fading memory properties of the dynamic model should be considered. In the context of theory-driven rational expectation models, Monte Carlo evidence of the behavior of SNPII estimators should be obtained in conjunction with alternative *solution methods* that deliver appropriate approximation of policy functions.

Finally, the benefits of SNPII estimation should also be analyzed in terms of its ability to deliver potentially better results in terms of *(i)* describing the nonlinear and asymmetric relation between economic variables, as thoroughly described in Granger and Terasvirta (1993); *(ii)* providing a more accurate account of the dynamic properties of data; and *(iii)* improving in-sample fit and out-of-sample forecasts.

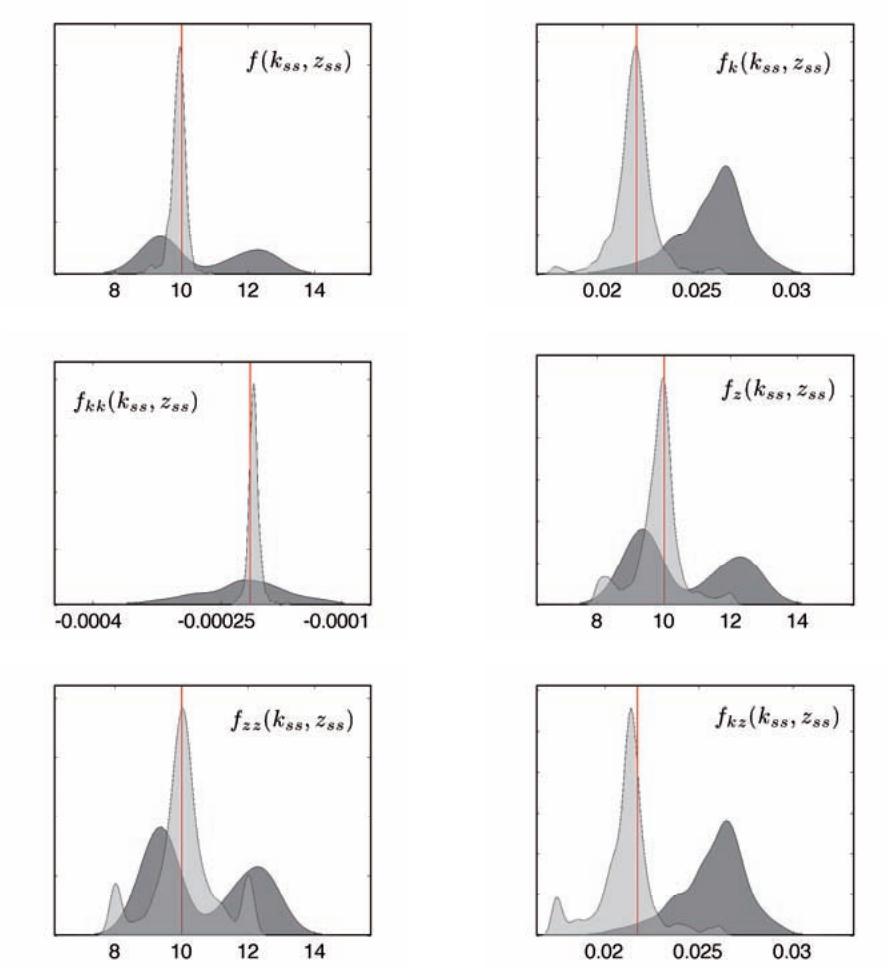

Figure 4.16: Density plots of estimated derivatives of f under SNPII estimator (light grey) and parametric II estimator (dark grey). True values of derivatives are marked by red vertical lines. $f(k_{ss}, z_{ss})$ corresponds to output level at steady-state. $f_k(k_{ss}, z_{ss})$ denotes the first derivative of estimated f w.r.t. capital at steady-state (i.e. the steady-state marginal productivity of capital). f_{kk} denotes the estimated second derivative w.r.t. capital. Likewise f_z, f_{zz} denote the estimated first and second derivatives of f w.r.t. TFP. Finally f_{kz} denotes the estimated cross partial derivative.

Chapter 5

Identifiable Uniqueness Conditions for a Large Class of Extremum Estimators

Proofs of consistency of extremum estimators usually require assumptions ensuring that there exists a unique well separated (*identifiably unique*) minimizer of the limit criterion function. Unfortunately, these assumptions are sometimes opaque and do not lend themselves to immediate verification. This is undesirable especially in the context of the SNPII methodology where it is important to establish easily the consistency of various auxiliary estimators of misspecified models.

This chapter provides methods for confirming that *identifiable uniqueness* holds for the class of extremum estimators whose limiting criterion function can be appropriately defined as a divergence on a space of probability measures (minimum distance estimators being a special case). In particular, it is shown that the task of verifying that *identifiable uniqueness* holds can be reduced to that of verifying the *strong unicity* of best approximations on an appropriate space of probability measures or regression functions. Some applications suggest that sufficient conditions for *strong unicity* of best approximations are often easy to verify, thus confirming the practical relevance of these methods.

5.1 Introduction

Building on early work of Doob (1934, 1953), Cramer (1946), Wald (1949), Le Cam (1953) and others that addressed the consistency of maximum likelihood (ML) estimators with independently identically distributed (iid) data, the "classic" consistency proof of extremum estimators originated in the well known contributions of Jennrich (1969) and Malinvaud (1970). These two papers independently addressed

the consistency of the least squares estimator in a nonlinear regression framework. They also seem to be at the origin of much of the research on the asymptotic properties of extremum estimators that took place during the following decades. Numerous contributions have since then allowed for the properties of extremum estimators to be well understood in multivariate dynamic settings, misspecified models and under heterogeneity and dependence of the data. The list is extensive. See e.g. Burguete et al. (1982), Amemiya (1983) and Gallant and White (1988b) for early reviews of important contributions, as well as Pötscher and Prucha (1991a,b, 1997) for a more recent and complete account of the relevant literature.

Despite the diversity, there is an underlying basic structure of conditions and methodologies that are common to the great majority of consistency results in this literature. In particular, the *uniform convergence* of criterion functions and the *identifiable uniqueness* of the argument that minimizes the limit criterion function seem to have pervasive influence, being present under many guises in most consistency proofs. Here we shall be concerned with the latter of these two conditions, the identifiable uniqueness, which requires fundamentally that the extremum estimator's limit criterion function have a well separated minimum (see e.g. White (1980a) and Domowitz and White (1982)).

Unfortunately, identifiable uniqueness conditions are sometimes opaque, in the sense that they do not seem to lend themselves to immediate verification. The suspicion of failure therefore remains; see e.g. Pötscher and Prucha (1991a, ch.4) for a review of problematic non-trivial cases where identifiable uniqueness fails to hold.

The aim of this chapter is to lay down a simple yet general methodology that allows the researcher to verify if the identifiable uniqueness assumption holds true in the context of possibly misspecified models. To follow the tradition of the "classic" results mentioned above we shall also adopt the nonlinear regression framework. For clarity, we consider here a simple prototypical nonlinear regression case that abstracts from the tedious considerations required by a more general result. It will become however clear that extensions to more general cases are often straightforward to achieve. Some trivial extensions to "non-regression" problems are mentioned here.

It should also be made clear from the outset that there is not necessarily a strict relation between imposing an identifiable uniqueness condition and ensuring that the model at hand satisfies the well known *identification condition* (even though this is often the case).[1] We do not address the identification condition here, although we discuss the role it plays in the present problem. An important practical implication of this distinction is that the present theory is mostly uninteresting for those spe-

[1]The researcher can always construct an extremum estimator (albeit possibly an uninteresting one) that satisfies an identifiable uniqueness condition despite having a model at hand that does not satisfy the fundamental identification condition, and vice-versa.

cial cases (typically involving well-specified models, compact parameter spaces and continuous criterion functions) where model identification implies that identifiable uniqueness holds on the estimator's criterion function.

Finally, it is also important to stress that we will be concerned with providing only sufficient conditions for identifiable uniqueness. Necessity is not addressed here. As such, the conditions under which the methodology remains of practical interest should be as general as possible. Indeed, it is not hard to devise restrictive conditions that once verified, imply immediately identifiable uniqueness (think e.g. of strict convexity of a continuous limit criterion function on a compact domain). Such conditions are however of very limited applicability and become, in that sense, uninteresting. The challenge is thus to achieve generality while at the same time ensuring simple verification.

The lack of a general enough methodology allowing researchers to verify if identifiable uniqueness assumptions hold has lead some authors to discuss the adequacy of this assumption in the context of misspecified models and to propose consistency results that do not rely on it; see e.g. Pötscher and Prucha (1991a, section 4.6) and references therein. We shall not follow this trail here.[2] We choose to follow instead the literature aimed at the verification of uniqueness conditions. Some examples include: *(i)* Freedman and Diaconis (1982) analyze inconsistency of redescending M-estimators for location parameters of symmetric distributions using iid data that is caused by failure of the uniqueness assumption; *(ii)* Kabaila (1983) addresses the failure of the uniqueness assumption for estimators of the parameter vector minimizing the one-step-ahead prediction errors in misspecified ARMA models; *(iii)* Clarke (1983) provides verifiable conditions for the uniqueness of ψ-type M-estimators using iid data that rely on somewhat restrictive conditions involving the Frechet differentiability of functional solutions; *(iv)* Rivest (1989) constitutes a failed attempt to prove uniqueness of robust extremum estimators, see Crisp and Burridge (1993); *(v)* Ducharme (1995) shows that the L_1-norm minimizer extremum estimator is generally unique in the context of well specified multivariate response nonlinear-regression models; *(vi)* Donoho and Liu (1988) observe pathologies of minimum-distance estimators related to the failure of uniqueness conditions (these pathologies can be well understood under the general methodology proposed here); and finally; *(vii)* Kent and Tyler (2001) provide conditions for local uniqueness of constrained and redescending M-estimators in the context of well-specified models, by imposing conditions for the estimator's criterion function to be locally well behaved.

[2]It often seems desirable to retain the identifiable uniqueness assumption as it provides the researcher with a host of useful properties, e.g. continuous mapping theorems for arg max functionals. Furthermore, this condition seems to play an important role in guaranteeing the economic interpretation of empirical work.

In what follows we generalize some of the results just mentioned in that we provide conditions for identifiable uniqueness to hold globally and for a large class of extremum estimators in the context of possibly dependent heterogeneous data and misspecified nonlinear regression models. In particular, we note that typical identifiable uniqueness assumptions can be restated in terms of transparent verifiable conditions on the nature of both the estimator and the model at hand. The idea is to adapt the statistical problem to be amenable to the use of results stemming from the field of Approximation Theory. These results are applicable to the class of extremum estimators whose limiting criterion function can be defined as a divergence on a space probability measures underlying the data. This class includes as a subset the usual minimum distance estimators. The problem of divergence minimization can also be translated to the space of regression functions. Building on Approximation Theory's results, the task of verifying the identifiable uniqueness of the limit minimizer is then reduced to that of verifying the strong uniqueness of best approximations in the space of probability measures or regression functions. Sufficient conditions for strong unicity are often easy to verify, thus giving the researcher the opportunity to check if identifiable uniqueness holds in various applications.

The rest of the chapter is structured as follows. Section 5.2 contains mainly preliminary considerations and lays down the foundations for the remaining sections both in terms of definitions and notation. Section 5.3 describes briefly the typical framework under which consistency of extremum estimators is obtained. Section 5.4 restates the estimation exercise in a more useful way by rewriting the limiting estimation problem as that of divergence minimization on the space of probability measures or regression functions. Section 5.5 introduces some concepts from Approximation Theory and reviews relevant results in this field highlighting the conditions under which approximation problems have (strongly) unique solutions. Section 5.6 derives identifiable uniqueness from this new set of conditions and provides some consistency results that follow immediately as corollaries. Section 5.7 illustrates the verification step with a few simple examples of nonlinear regression models and alternative extremum estimators. Section 5.8 concludes.

Finally, a word on notation. In what follows, $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{R}$ denote the sets of natural, integer and real numbers. If $\mathcal{A}$ is a set, $\mathfrak{B}(\mathcal{A})$ denotes the Borel σ-algebra over $\mathcal{A}$, and $\times_{t=1}^{t=T}\mathcal{A}$, often denoted $\mathcal{A}_T$, is the Cartesian product of T copies of $\mathcal{A}$. Furthermore, in linear spaces, boldfaced letters (e.g. $\mathbf{a} \in \mathcal{A}$) denote vectors. Note also that $:=$ denotes *definitional equivalence*, whereas $\equiv$ is used to denote *practical equivalence*. If f and g are maps, then $f \circ g := f(g)$ denotes their composition. The mappings $d_{\mathcal{A}}$ and $d^*_{\mathcal{A}}$ denote a divergence and metric defined on the set $\mathcal{A} \times \mathcal{A}$ respectively, and $\|\cdot\|_{\mathcal{A}}$ denotes a norm on $\mathcal{A}$. Finally, p.m. and a.s. stand for *probability measure* and *almost surely*, respectively.

5.2 Preliminary Considerations

This section is sometimes dense and the casual reader might prefer to use it exclusively as a reference for notation and definitions, thus proceeding directly to Section 5.3. Consider the T-period sequence $\{\mathbf{x}_t(\omega)\}_{t=1}^{T}$, a subset of the realized path of an n_x-variate stochastic sequence $\mathbf{x}(\omega) := \{\mathbf{x}_t(\omega), t \in \mathbb{Z}\}$, for some $\omega \in \Omega$ the event space. Let $\mathbf{x}_t(\omega) \in \mathcal{X} \subseteq \mathbb{R}^{n_x}\ \forall (\omega, t) \in \Omega \times \mathbb{Z}$.[3] The random sequence $\mathbf{x}$ is thus an $\mathcal{F}/\mathfrak{B}(\mathcal{X}_\infty)$-measurable mapping $\mathbf{x} : \Omega \to \mathcal{X}_\infty \subseteq \mathbb{R}_\infty^{n_x}$ where $\mathbb{R}_\infty^{n_x} := \times_{t=-\infty}^{t=\infty}\mathbb{R}^{n_x}$ denotes the Cartesian product of infinite copies of $\mathbb{R}^{n_x}$ and $\mathcal{X}_\infty = \times_{t=-\infty}^{t=\infty}\mathcal{X}$ with $\mathfrak{B}(\mathcal{X}_\infty) \equiv \mathfrak{B}(\mathbb{R}_\infty^{n_x}) \cap \mathcal{X}_\infty$ (Billingsley (1995, p.159)) where $\mathfrak{B}(\mathbb{R}_\infty^{n_x})$ denotes the Borel σ-algebra generated by the finite dimensional product cylinders of $\mathbb{R}_\infty^{n_x}$, $\mathcal{F}$ denotes a σ-field defined on the event space Ω, and together with the p.m. P_0 on $\mathcal{F}$, the triplet $(\Omega, \mathcal{F}, \mathrm{P}_0)$ denotes the complete probability space of interest. For every $\omega \in \Omega$, the stochastic sequence $\mathbf{x}(\omega)$ thus lives on the space $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), \mathrm{D}_0^{\mathbf{x}})$ where the p.m. $\mathrm{D}_0^{\mathbf{x}}$ is defined over the elements of $\mathfrak{B}(\mathcal{X}_\infty)$. Following White (1980b) and Domowitz and White (1982), consider now the univariate stochastic sequence,

$$y := \{y_t = h_0(\mathbf{x}_t), t \in \mathbb{Z}\}$$

with $h_0 : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ an $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable mapping, $\mathfrak{B}(\mathcal{X}) \equiv \mathcal{X} \cap \mathfrak{B}(\mathbb{R}^{n_x})$ and $\mathfrak{B}(\mathcal{Y}) \equiv \mathcal{Y} \cap \mathfrak{B}(\mathbb{R})$. The results in this chapter can be easily extended to more complex high dimensional nonlinear dynamic models with unobserved variables and possibly intractable likelihood functions.[4] Hence, for every $t \in \mathbb{Z}$, $h_0 \circ \mathbf{x}_t : \Omega \to \mathcal{Y}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{Y})$-measurable. For every $\omega \in \Omega$, the sequence $y(\omega)$ thus lives in the space $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), \mathrm{D}_0^y)$ where D_0^y is the p.m. induced by h_0 on $\mathfrak{B}(\mathcal{Y}_\infty)$ according to $\mathrm{D}_0^y(B_y) = \mathrm{D}_0^{\mathbf{x}} \circ h_0^{-1}(B_y)\ \forall B_y \in \mathfrak{B}(\mathcal{Y}_\infty)$. Define now the joint process $\mathbf{w} := \{\mathbf{w}_t = (y_t, \mathbf{x}_t), t \in \mathbb{Z}\}$. For every $\omega \in \Omega$, $\mathbf{w}_t(\omega) \in \mathcal{W} \equiv \mathcal{Y} \times \mathcal{X}$ and $\mathbf{w}(\omega) \in \mathcal{W}_\infty \equiv \mathcal{Y}_\infty \times \mathcal{X}_\infty \subseteq \mathbb{R}_\infty^{1+n_x} \equiv \times_{t=-\infty}^{t=\infty}\mathbb{R}^{1+n_x}$. The sequence thus lives in $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), \mathrm{D}_0^{\mathbf{w}})$ where $\mathrm{D}_0^{\mathbf{w}}$ denotes the measure defined on $\mathfrak{B}(\mathcal{W}_\infty) \equiv \mathcal{W}_\infty \cap \mathfrak{B}(\mathbb{R}_\infty^{1+n_x})$.[5] Finally, suppose that for some $\omega \in \Omega$ the T-period sequence $\mathbf{w}_\mathbf{T}(\omega) := (\mathbf{y}_\mathbf{T}(\omega), \mathbf{x}_\mathbf{T}(\omega))$ is observed, where $y_\mathbf{T}(\omega) := \{y_t(\omega)\}_{t=1}^{t=T}$ and $\mathbf{x}_\mathbf{T}(\omega) := \{\mathbf{x}_t(\omega)\}_{t=1}^{t=T}$. Yet, h_0 is unknown.

A postulated parametric regression model takes the form $\hat{y}_t = h(\mathbf{x}_t; \boldsymbol{\theta})$ so that the modeled counterpart of the stochastic sequence y is given by,

$$\hat{y} := \{\hat{y}_t = h(\mathbf{x}_t; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}$$

[3] Properties of the data in terms of dynamics and heterogeneity are addressed in Section 5.3.

[4] Thus the extension covers what is probably the most common formulation of the nonlinear regression $y_t = h_0(\mathbf{x}_t) + \epsilon_t$ where ϵ_t is unobserved. Here we follow White (1980b) in considering an extremely simple univariate nonlinear regression framework. This allows us to simplify the argument by focusing on what is really essential, therefore avoiding distractions created by unnecessary considerations.

[5] $\mathfrak{B}(\mathcal{W}_\infty) = \mathfrak{B}(\mathcal{X}_\infty) \otimes \mathfrak{B}(\mathcal{Y}_\infty)$ the product σ-algebra; Dudley (2002, p.119).

where $h : \mathcal{X} \times \Theta \to \mathcal{Y}$. Here we deviate slightly from standard notation. The use of the hat over y does not imply that fitted values are obtained at a specific point of Θ (usually some $\hat{\boldsymbol{\theta}}_T(\omega)$, $\omega \in \Omega$). In the present context, the hat is used only to distinguish *modeled data* from *observed data*. Also, we allow Θ to be infinite dimensional (although typically metrizable). By *parametric model* we just mean a set of p.m.s that is indexed by a parameter $\boldsymbol{\theta} \in \Theta$. In this sense, we also deviate somewhat from typical terminology that requires Θ to be finite dimensional. For every $\boldsymbol{\theta} \in \Theta$, let $h(\cdot, \boldsymbol{\theta}) : \mathcal{X} \to \mathcal{Y}$ be $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$-measurable, so that $h(\mathbf{x}_t; \boldsymbol{\theta}) : \Omega \to \mathcal{Y}$ is $\mathcal{F}/\mathfrak{B}(\mathcal{Y})$-measurable $\forall \boldsymbol{\theta} \in \Theta$ and every $t \in \mathbb{Z}$. Define $\mathcal{H}_\Theta(\mathcal{X}) := \{h(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ as the space of parametric functions defined on $\mathcal{X}$ generated by Θ under the mapping $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ where $h_\mathcal{X}(\boldsymbol{\theta}) := h(\cdot; \boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \, \forall \boldsymbol{\theta} \in \Theta$. The mapping $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ shall be called a *parameterization mapping*. Immediately, given $\mathrm{D}_0^{\mathbf{x}}$, for every $\boldsymbol{\theta} \in \Theta$, $h(\cdot, \boldsymbol{\theta}) \equiv h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ induces a p.m. $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}}$ indexed by $\boldsymbol{\theta}$ on $\mathfrak{B}(\mathcal{Y}_\infty)$ according to $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}}(B_y) = \mathrm{D}_0^{\mathbf{x}} \circ h^{-1}(B_y, \boldsymbol{\theta})$ for every $(B_y, \boldsymbol{\theta}) \in \mathfrak{B}(\mathcal{Y}_\infty) \times \Theta$. The triplet $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), \mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}})$ is thus an element of a family of measure spaces indexed by $\boldsymbol{\theta}$. Now, define accordingly $\hat{\mathbf{w}} := \{\hat{\mathbf{w}}_t = (\hat{y}_t, \mathbf{x}_t), t \in \mathbb{Z}\}$, the counterpart of $\mathbf{w}$, with $\hat{\mathbf{w}}_t(\omega) \in \mathcal{W} \, \forall t \in \mathbb{N}$ and $\hat{\mathbf{w}}(\omega) \in \mathcal{W}_\infty$ which lives in $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$. As a result, given $\mathrm{D}_0^{\mathbf{x}}$, for every $\boldsymbol{\theta} \in \Theta$, $h_\mathcal{X}(\boldsymbol{\theta})$ induces also a p.m. $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ on $\mathfrak{B}(\mathcal{W}_\infty)$ so that $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ is also indexed by $\boldsymbol{\theta}$. For clarity, we let D denote the functional that, given $\mathrm{D}_0^{\mathbf{x}}$ on $\mathfrak{B}(\mathcal{X}_\infty)$, maps elements of $\mathcal{H}_\Theta(\mathcal{X})$ onto the space $\mathcal{D}_\Theta^{\hat{\mathbf{w}}}$ of p.m.s defined on the sets of $\mathfrak{B}(\mathcal{W}_\infty)$ and generated by Θ through h, i.e. $\mathrm{D} : \mathcal{H}_\Theta(\mathcal{X}) \to \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$ (with $\mathcal{D}_\Theta^{\hat{\mathbf{w}}} = \{\mathrm{D} \circ h_\mathcal{X}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \mid \mathrm{D}_0^{\mathbf{x}}\}$) satisfies $\mathrm{D} \circ h_\mathcal{X}(\boldsymbol{\theta}) = \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \, \forall \boldsymbol{\theta} \in \Theta$ with $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(B_{\mathbf{w}}) \equiv \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(B_{\mathbf{x}} \times B_y) = \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(B_{\mathbf{x}} \times \mathcal{Y}_\infty | \mathcal{X}_\infty \times B_y) \cdot \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(\mathcal{X}_\infty \times B_y) = I_{(B_{\mathbf{x}} = h^{-1}(B_y))} \cdot \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}(\mathcal{X}_\infty \times B_y)$, $B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X})$ and $B_y \in \mathfrak{B}(\mathcal{X})$ with $I_{(B_{\mathbf{x}} = h^{-1}(B_y))} = 1$ when $B_{\mathbf{x}} = h^{-1}(B_y)$ and $I_{(B_{\mathbf{x}} = h^{-1}(B_y))} = 0$ otherwise.[6] Clearly, since there is no guarantee that $h_0 \in \mathcal{H}_\Theta(\mathcal{X})$, i.e. that $\exists \boldsymbol{\theta}_0 \in \Theta : h(\mathbf{x}_t(\omega); \boldsymbol{\theta}_0) = h_0(\mathbf{x}_t(\omega)) \, \forall \mathbf{x}_t(\omega) \in \mathcal{X}$, it might well be the case that $\nexists \boldsymbol{\theta}_0 \in \Theta : \mathrm{D} \circ h_\mathcal{X}(\boldsymbol{\theta}_0) = \mathrm{D}_0^{\mathbf{w}}$ so that $\mathrm{D}_0^{\mathbf{w}} \notin \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$. Note here that the statement $\exists \boldsymbol{\theta}_0 \in \Theta : h(\mathbf{x}_t; \boldsymbol{\theta}_0) = h_0(\mathbf{x}_t) \, \forall \mathbf{x}_t \in \mathcal{X}$ is to be understood in the function equivalence sense (Kolmogorov and Fomin (1975), p.288); i.e. we write $h_\mathcal{X}(\boldsymbol{\theta}_0) = h_0$ if and only if $\mathrm{D}_0^{\mathbf{x}}\{B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty) : h_0(B_{\mathbf{x}}) \neq h(B_{\mathbf{x}}; \boldsymbol{\theta})\} \equiv P\{\omega \in \Omega : h_0(\mathbf{x}(\omega)) \neq h(\mathbf{x}(\omega); \boldsymbol{\theta})\} = 0$. The same applies to similar statements throughout this chapter. The sets $\mathcal{H}_\Theta(\mathcal{X})$ and Θ are thus naturally partitioned into equivalence classes by the mappings D and $h_\mathcal{X}$ respectively, with classes taking the form $\{h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) : \mathrm{D} \circ h_\mathcal{X}(\boldsymbol{\theta}) = \mathrm{D} \circ h_\mathcal{X}(\boldsymbol{\theta}')\}$ and $\{\boldsymbol{\theta} \in \Theta : h_\mathcal{X}(\boldsymbol{\theta}) = h_\mathcal{X}(\boldsymbol{\theta}')\}$ respectively. This framework is convenient as the identification problem is not the

[6] By "given $\mathrm{D}_0^{\mathbf{x}}$" we mean that $\mathrm{D} : \mathcal{H}_\Theta(\mathcal{X}) \to \mathrm{D}_\Theta^{\hat{\mathbf{w}}}$ can be obtained from $\mathrm{D}^* : \mathcal{D}^{\mathbf{x}} \times \mathcal{H}_\Theta(\mathcal{X}) \to \mathrm{D}_\Theta^{\hat{\mathbf{w}}}$ as $\mathrm{D} = \mathrm{D}^*(\mathrm{D}_0^{\mathbf{x}}, \cdot) : \mathcal{H}_\Theta(\mathcal{X}) \to \mathrm{D}_\Theta^{\hat{\mathbf{w}}}$ where $\mathrm{D}_0^{\mathbf{x}} \in \mathcal{D}^{\mathbf{x}}$. Also note that every $B_{\mathbf{w}} \in \mathfrak{B}(\mathcal{W}_\infty)$ takes the form $B_{\mathbf{w}} = B_{\mathbf{x}} \times B_y$ with $B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty)$ and $B_y \in \mathfrak{B}(\mathcal{Y}_\infty)$ (Dudley (2002, p.118)); see also Davidson (1994, p.115) for notation.

one we wish to focus on. We shall address this point later. Finally, let,

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_{\mathbf{T}}; \boldsymbol{\theta})$$

denote the extremum estimator of interest, a map $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$. For the moment, let us adopt this notation to stress that Q_T is a function of $\boldsymbol{\theta} \in \Theta$. Hence, we write $Q_T : \mathcal{W}_T \times \Theta \to \mathbb{R}$ where $\mathcal{W}_T := \mathcal{Y}_T \times \mathcal{X}_T$ with $\mathcal{Y}_T := \times_{t=1}^{t=T}\mathcal{Y}$ and $\mathcal{X}_T := \times_{t=1}^{t=T}\mathcal{X}$ so that $\mathbf{w}_{\mathbf{T}}(\omega) \in \mathcal{W}_T$. Note however that we could have written $Q_T(\mathbf{x}_{\mathbf{T}}, \mathbf{h}_0(\mathbf{x}_{\mathbf{T}}), \mathbf{h}(\mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta}))$ where $\mathbf{h}_0(\mathbf{x}_{\mathbf{T}}) := \{h_0(\mathbf{x}_t)\}_{t=1}^T \equiv y_{\mathbf{T}}$ and $\mathbf{h}(\mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta}) := \{h(\mathbf{x}_t; \boldsymbol{\theta})\}_{t=1}^T \equiv \hat{y}_{\mathbf{T}}$ to highlight the fact that the criterion Q_T is a function of $\boldsymbol{\theta}$ through $h_{\mathcal{X}}$, and as a result, that $\hat{\boldsymbol{\theta}}_T$ depends also on the choice of parameterization. For simplicity however, since $h_{\mathcal{X}}$ is often fixed prior to estimation, an explicit account of this relation is seldom considered. Clearly, nothing is lost in adopting either notational convention as long as these considerations are kept in mind.

Finally, note that we can also address the problem of approximating the true distribution D_0^y of a random variable y_t from a family of parametric distributions $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}}$, simply by taking $\mathrm{D}_0^{\mathbf{x}}$ to be known. For example, taking $\mathbf{x}$ to be independently identically distributed, with $n_x = 1$ and $x_t \sim \mathcal{U}([0,1])$ where $\mathcal{U}([0,1])$ denotes the uniform distribution on $[0,1]$, implies that $\mathrm{D}_0^y = h_0^{-1}$ is the true unknown distribution of y_t and that $h_{\mathcal{X}}^{-1}(\boldsymbol{\theta})$ defines the distribution function $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{y}} = h^{-1}(\cdot; \boldsymbol{\theta})$ of $\hat{y}_t$. Also note that the results in this chapter extend trivially to a formulation of the regression model where $y_t = h_0(\mathbf{x}_t) + \epsilon_t$ whenever the distribution of ϵ_t is known, or more generally to $y_t = H(h_0(\mathbf{x}_t), \epsilon_t)$, $\epsilon_t \sim F_\epsilon$ whenever H and F_ϵ are known.

5.3 Standard Formulation

Following White (1980b) and Domowitz and White (1982), consider for simplicity the regression model $y_t = h_0(\mathbf{x}_t)$ and a postulated parametric counterpart $\hat{y}_t = h(\mathbf{x}_t, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Existence of an estimator $\hat{\boldsymbol{\theta}}_T$ as described above follows immediately from lemma 2 of Jennrich (1969) and Pötscher and Prucha (1997, p.20, lemma 3.4); see also e.g. Brown and Purves (1973) and Stinchcombe and White (1992) for generalizations and extensions.

Assumption 5.3.1. *Θ is compact and $Q_T(\mathbf{w}_{\mathbf{T}}(\omega); \cdot) : \Theta \to \mathbb{R}$ is a continuous function of $\boldsymbol{\theta} \in \Theta$ for every $\mathbf{w}_{\mathbf{T}}(\omega) \in \mathcal{W}_T$, (i.e. every $\omega \in \Omega$). Also, $Q_T(\cdot; \boldsymbol{\theta}) : \mathcal{W}_T \to \mathbb{R}$ is a $\mathfrak{B}(\mathcal{W}_T)/\mathfrak{B}(\mathbb{R})$-measurable function of $\mathbf{w}_{\mathbf{T}}$ for every $\boldsymbol{\theta} \in \Theta$.*

Lemma 5.3.1. (Existence) *Let Assumption 5.3.1 hold. Then there exists a measurable function $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ such that for every $\omega \in \Omega$ we have $Q_T(\mathbf{w}_{\mathbf{T}}(\omega); \hat{\boldsymbol{\theta}}_T(\omega)) = \min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w}_{\mathbf{T}}(\omega); \boldsymbol{\theta})$.*[7]

[7] Assumption 5.3.1 and Lemma 5.3.1 can be further generalized to accommodate cases under

Consistency of $\hat{\boldsymbol{\theta}}_T$ has been obtained under conditions that ensure *(i)* the convergence of the sequence of continuous functions $Q_T : \mathcal{W}_T \times \Theta \to \mathbb{R}$ as $T \to \infty$, to a limit deterministic function $Q_\infty : \Theta \to \mathbb{R}$, uniformly on Θ, and *(ii)* the identifiable uniqueness of $\boldsymbol{\theta}_0 := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$. Definition 5.3.1 is adapted from Bates and White (1985).[8]

Definition 5.3.1. (Identifiable Uniqueness) *Suppose that $\boldsymbol{\theta}_0$ minimizes Q_∞ on Θ. Let $S_0(\epsilon)$ be an open ball centered at $\boldsymbol{\theta}_0$ with radius $\epsilon > 0$. Define the neighborhood $\eta_0(\epsilon) \equiv S_0(\epsilon) \subset \Theta$ with complement $\eta_0(\epsilon)^c := \Theta \backslash \eta_0(\epsilon)$. Then $\boldsymbol{\theta}_0$ is said to be identifiable unique on Θ iff for every $\epsilon > 0$, $\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0$.*

In general, the identifiable uniqueness of $\boldsymbol{\theta}_0$ allows for alternative formulations of consistency of extremum estimators in terms of non-compact parameter spaces, discontinuous criterion functions, as well as for dependence and heterogeneity of the underlying data. In particular, this condition can be formulated for sequences of minimizers $\boldsymbol{\theta}_0^T$ of a sequence of deterministic functions Q_∞^T to which the random criterion function Q_T converges. For the sake of simplicity however, we shall ignore this possibility. We thus focus only on the case where $Q_\infty^T \equiv Q_\infty \; \forall T$. Lemma 5.3.2 below is adapted from Pötscher and Prucha (1997, ch.3).

Assumption 5.3.2. $\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\mathbf{w_T}; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \overset{a.s.}{\to} 0$.

Assumption 5.3.3. *$Q_\infty : \Theta \to \mathbb{R}$ has an identifiably unique minimizer $\boldsymbol{\theta}_0$.*

Lemma 5.3.2. (Consistency) *Let Assumptions 5.3.2 and 5.3.3 hold. Define $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ such that $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}; \boldsymbol{\theta})$. Then, $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \overset{a.s.}{\to} 0$ as $T \to \infty$.*

Assumption 5.3.1 can be added to Assumptions 5.3.2 and 5.3.3 in the lemma above to ensure that $\hat{\boldsymbol{\theta}}_T$ is a random variable for every T. This however, is not a necessary condition for the measurability of $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$, nor is it necessary to obtain $\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0 \overset{a.s.}{\to} 0$ as the lemma itself testifies. Still, when it is appropriate to work under the influence of Assumption 5.3.1, then, given the compactness of Θ and the continuity of Q_∞, the identifiable uniqueness condition turns out to be satisfied as long as the set $\arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta})$ is a singleton, i.e. $\boldsymbol{\theta}_0$ is unique. Sometimes, it will be perfectly fine to consider the set of elementary Assumptions 5.3.1, 5.3.2 and 5.3.4 (below), and to work with the following lemma adapted from Amemiya (1985).

Assumption 5.3.4. *$Q_\infty : \Theta \to \mathbb{R}$ attains a unique minimum at $\boldsymbol{\theta}_0$.*

which Q is continuous on Θ a.s. but not necessarily for all $\omega \in \Omega$; see e.g. Gallant and White (1988b, p.14).

[8] The uniform convergence condition is typically stronger than required; see e.g. van der Vaart and Wellner (1996, p.286) and Pötscher and Prucha (1997, p.24).

Lemma 5.3.3. (Consistency) *Let Assumptions 5.3.1, 5.3.2 and 5.3.4 hold. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}; \boldsymbol{\theta})$. *Then* $\hat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}_0$ *as* $T \to \infty$.

Finally, a few comments on Assumptions 5.3.2-5.3.4. Well known conditions for $\sup_{\theta\in\Theta} |Q_T(\cdot;\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \to 0$ a.s. on a totally bounded metric space Θ are *(i)* $Q_T(\cdot;\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}) \to 0$ a.s. pointwise for every $\boldsymbol{\theta} \in \Theta$ and *(ii)* $\{Q_T(\cdot,\boldsymbol{\theta}), T \in \mathbb{N}\}$ be strongly asymptotically uniformly stochastically equicontinuous (see e.g. Newey (1991) and Andrews (1992)). When $\{Q_T(\cdot,\boldsymbol{\theta}), T \in \mathbb{N}\}$ is a sequence of normalized partial sums, Assumption 5.3.2 boils down to a uniform law of large numbers. These have been achieved under alternative primitive conditions that allow for varying degrees of dependence and heterogeneity in the data; see e.g. Gallant and White (1988b, ch.3) and Pötscher and Prucha (1997, ch.5) and references therein.[9]

Statistical tests have been developed that are aimed at verifying whether (at least a part of) the host of assumptions involved in these arguments actually hold in practice. To some extent, this allows researchers to conclude with varying degree of confidence on whether the consistency of any given extremum estimator holds. Unfortunately, in the context of misspecified models, it is often hard to conclude wether the identifiable uniqueness assumption is satisfied. As mentioned in the introduction, some authors have attempted (often successfully) to relax this condition and allow for multiple minima. This might be a fruitful approach in some circumstances, albeit one that we shall not follow here. Below we investigate transparent primitive conditions on both the estimator and the model at hand that imply identifiable uniqueness. These conditions take place in a deterministic setting as they pertain to the limit criterion function. The uniform convergence of the criterion function, established in a probabilistic setting, will be left unaltered.

5.4 Limit Divergence Functions

As mentioned before, the functional dependence of Q_T on the choice of parameterization mapping $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is typically omitted for notational convenience. Recall from Section 5.2 that we could have written $Q_T(\mathbf{x_T}, \mathbf{h}_0(\mathbf{x_T}), \mathbf{h}(\mathbf{x_T};\boldsymbol{\theta})) \equiv Q_T(\mathbf{w_T}, \hat{\mathbf{w}}_\mathbf{T}(\boldsymbol{\theta}))$ thus having $\hat{\boldsymbol{\theta}}_T = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{w_T}, \hat{\mathbf{w}}_\mathbf{T}(\boldsymbol{\theta}))$. This clarifies the reason why the deterministic limit criterion is often appropriately described as a function $Q_\infty^\mathcal{D}$ of the underlying joint p.m.s of $\mathbf{w_T}$ and $\hat{\mathbf{w}}_\mathbf{T}$ (or some of its features, e.g. moments) implicitly defined by the measurable mappings h_0 and $h_\mathcal{X}(\boldsymbol{\theta})\ \forall \boldsymbol{\theta} \in \Theta$, given $\mathrm{D}_0^\mathbf{x}$. Below, we shall restrict attention to limit criterion functions $Q_\infty : \Theta \to \mathbb{R}$

[9] When $Q_T(\cdot;\boldsymbol{\theta}) \equiv T^{-1}\sum_{t=1}^T q(\mathbf{w}_t;\boldsymbol{\theta})$ uniform convergence is equivalent to $\mathcal{Q} = \{q(\cdot;\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ being a class of Glivenko-Cantelli functions. This requires fundamentally the compactness of Θ, continuity of $q(\mathbf{w_T};\cdot) : \Theta \to \mathbb{R}$ for every $\mathbf{w_T} \in \mathcal{W}_T$ (i.e. every $\omega \in \Omega$) and that $q(\cdot;\boldsymbol{\theta})$ be dominated by an integrable function for every $\boldsymbol{\theta} \in \Theta$.

that assume the special form $Q_\infty(\boldsymbol{\theta}) = Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})\ \forall \boldsymbol{\theta} \in \Theta$ where $\mathrm{D}_0^{\mathbf{w}}$ and $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ are the p.m.s of the processes $\mathbf{w}$ and $\hat{\mathbf{w}}$ defined in Section 5.2. When $Q_\infty^{\mathcal{D}}$ is a divergence on a space of probability measures containing $\mathrm{D}_0^{\mathbf{w}}$ and $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}\ \forall \boldsymbol{\theta} \in \Theta$, then $\boldsymbol{\theta}_0$ is, by definition, the minimizer of that divergence.[10] By establishing a bijection between the space of probability measures and the space of regression functions containing h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})\ \forall \boldsymbol{\theta} \in \Theta$, we translate the problem of divergence minimization from the space distributions to the space of regression functions. Finally, we also note that it is possible to simplify the argument when there exists a strictly increasing function g such that $g \circ Q_\infty^{\mathcal{D}}$ establishes a metric, norm or inner product on the space of distributions (and regression functions), in which case $\boldsymbol{\theta}_0$ is characterized as the minimizer of this distance between $\mathrm{D}_0^{\mathbf{w}}$ and $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ (or h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$). In Section 5.5, we review conditions for (strong) uniqueness of best approximations of $\mathrm{D}_0^{\mathbf{w}}$ by $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$ in the space of probability measures (or h_0 by $h_{\mathcal{X}}(\boldsymbol{\theta})$ in the space of regression functions).

Consider the space $\mathcal{H}(\mathcal{X})$ satisfying $\mathcal{H}_\Theta(\mathcal{X}) \subseteq \mathcal{H}(\mathcal{X})$ and $h_0 \in \mathcal{H}(\mathcal{X})$. The smallest $\mathcal{H}(\mathcal{X})$ thus being $\mathcal{H}(\mathcal{X}) = \mathcal{H}_\Theta(\mathcal{X}) \times \{h_0\}$ when $h_0 \notin \mathcal{H}_\Theta(\mathcal{X})$ or simply $\mathcal{H}(\mathcal{X}) = \mathcal{H}_\Theta(\mathcal{X})$ when $\exists \boldsymbol{\theta}_0 \in \Theta : h_{\mathcal{X}}(\boldsymbol{\theta}_0) = h_0$ which implies $h_0 \in \mathcal{H}_\Theta(\mathcal{X})$. Now, define the space of p.m.s $\mathcal{D}^{\mathbf{w}} = \{\mathrm{D}(h), h \in \mathcal{H}(\mathcal{X})\}$ by extending the functional D encountered before to be defined on $\mathcal{H}(\mathcal{X})$ instead of $\mathcal{H}_\Theta(\mathcal{X})$ only; i.e. now $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$, so that in general D is such that $\mathrm{D} \circ h = \mathrm{D}_h^{\mathbf{w}}$ with $\mathrm{D}_h^{\mathbf{w}}$ satisfying $\mathrm{D}_h^{\mathbf{w}}(B_{\mathbf{w}}) \equiv \mathrm{D}_h^{\mathbf{w}}(B_y, B_{\mathbf{x}}) \equiv \mathrm{D}_h^{y|\mathbf{x}}(B_y) \cdot \mathrm{D}_0^{\mathbf{x}}(B_{\mathbf{x}}) \equiv \mathrm{D}^{y|\mathbf{x}}(B_y|h) \cdot \mathrm{D}_0^{\mathbf{x}}(B_{\mathbf{x}})\ \forall (h, B_{\mathbf{w}}) \in \mathcal{H}(\mathcal{X}) \times \mathfrak{B}(\mathcal{W}_\infty)$. It thus follows that $\mathcal{D}^{\mathbf{w}}$ satisfies $\mathcal{D}_\Theta^{\hat{\mathbf{w}}} \subseteq \mathcal{D}^{\mathbf{w}}$ and $\mathrm{D}_0^{\mathbf{w}} \in \mathcal{D}(\mathcal{X})$. The smallest $\mathcal{D}^{\mathbf{w}}$ corresponding to the smallest $\mathcal{H}(\mathcal{X})$ and defined as $\mathcal{D}^{\mathbf{w}} = \mathcal{D}_\Theta^{\hat{\mathbf{w}}} \times \{\mathrm{D}_0^{\mathbf{w}}\}$ for misspecified models or simply $\mathcal{D}^{\mathbf{w}} = \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$ when the model is well specified, i.e. when $\exists \boldsymbol{\theta}_0 \in \Theta : \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}_0) = \mathrm{D} \circ h_0 = \mathrm{D}_0^{\mathbf{w}}$ (which implies $\mathrm{D}_0^{\mathbf{w}} \in \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$). Finally, let the following assumption restrict the class of extremum estimators under consideration.

Assumption 5.4.1. *The limit criterion $Q_\infty : \Theta \to \mathbb{R}$ takes the form $Q_\infty(\boldsymbol{\theta}) \equiv Q_\infty^{\mathcal{D}}(\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}, \mathrm{D}_0^{\mathbf{w}})\ \forall \boldsymbol{\theta} \in \Theta$ where $Q_\infty^{\mathcal{D}} : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a divergence $d_{\mathcal{D}} \equiv Q_\infty^{\mathcal{D}}$ on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$.*

Since under Assumption 5.4.1, $Q_\infty^{\mathcal{D}}$ is a function of $\boldsymbol{\theta} \in \Theta$ only through the p.m. $\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \equiv \mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{D}_\Theta^{\hat{\mathbf{w}}}$, we require that $\nexists (\boldsymbol{\theta}', \boldsymbol{\theta}'') \in \Theta \times \Theta$ satisfying $\boldsymbol{\theta}' \neq \boldsymbol{\theta}''$ and such that $\mathrm{D} \circ h_{\mathcal{X}}(\boldsymbol{\theta}') = \mathrm{D} \circ h(\cdot; \boldsymbol{\theta}'')$ as a minimal condition for uniqueness. In several contexts, this is called the *identification condition* (see e.g. Hsiao (1983)).

[10] Given a limit criterion function $Q_\infty : \Theta \to \mathbb{R}$ and a flexible definition of divergence (e.g. a pre-metric), it is often possible to find a divergence $Q_\infty^{\mathcal{D}}$ on the space of p.m.s satisfying $\arg\min_{\boldsymbol{\theta} \in \Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}, \mathrm{D}_0^{\mathbf{w}}) = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta})$. In this sense, the results discussed here are generally applicable to a large number of extremum estimator, even those not initially conceived as minimum divergence estimators.

As mentioned in the introduction, there is no universal strict relation between identifiable uniqueness and identification. In most cases of interest however, the absence of observationally equivalent elements in Θ is a necessary condition for identifiable uniqueness to hold. This is also the case in our formulation where the limit criterion $Q_\infty^{\mathcal{D}}$ takes the form of a divergence on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$.[11]

In the present context, for $\mathrm{D} \circ h_{\mathcal{X}} : \Theta \to \mathcal{D}^{\mathbf{w}}$ to be injective, a necessary and sufficient condition is that both $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ and $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ be one-to-one. Now, the injective nature of D is often unverifiable, since it is in the very nature of statistical inference that the true probability measure $\mathrm{D}_0^{\mathbf{w}}$ be not known. In simple cases, depending on the complexity of $\mathcal{H}(\mathcal{X})$, it might be possible to find convincing evidence that $\mathrm{D}_0^{\mathbf{x}}$ is rich enough for D to be injective, based on observed data alone.[12] Yet, this is not always the case and little can be done about it as long as $\mathrm{D}_0^{\mathbf{w}}$ is to remain unknown. There is thus no point in discussing this issue further and we proceed under the common assumption that the data is "rich enough" for different elements of $\mathcal{H}(\mathcal{X})$ to be identified as such.[13] Clearly, the researcher might feel more or less comfortable in imposing this assumption depending on the complexity of $\mathcal{H}(\mathcal{X})$ and on the evidence contained in observed data. Still, imposing some condition on the richness of the data seems simply unavoidable. As mentioned in Section 5.2, it is important to note that this assumption is already embodied in the function equivalence framework adopted here, so that $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ is bijective by construction.

It is thus evident that the one-to-one nature of the composition $\mathrm{D} \circ h_{\mathcal{X}} : \Theta \to \mathcal{D}^{\mathbf{w}}$ is to be understood fundamentally as a restriction on the construction of the model (in particular on the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$) as it does not concern the estimation procedure nor does it involve considerations about the data generating process beyond those already covered by the function equivalence framework adopted throughout this chapter. Also, note that since the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is surjective by construction, and $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ is bijective (also by construction), the only property of concern to us is that $h_{\mathcal{X}}$ be injective. This is generally verifiable for any given class of parametric functions posited by the researcher, and it is controlled by the researcher, so it should be satisfied by an appropriate formulation of the regression model and the parameter space Θ.[14] Still, we let the injective nature of $h_{\mathcal{X}}$ be sated as an assumption for future reference and verification.

[11] This would not be the case if the limit criterion was instead defined more generally on e.g. $\Omega \times \Theta$.

[12] Think e.g. of a simple linear regression with observed $\mathbf{w}_\mathbf{T}$ providing evidence of a rich $\mathrm{D}_0^{\mathbf{w}}$.

[13] A "rich" data setting should exclude e.g. the presence of degenerate and collinear-type random variables.

[14] As we shall see in Section 5.7, verification of Assumption 5.4.2 is often a straightforward exercise.

Assumption 5.4.2. *$h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is injective.*

This assumption implies by construction that both $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ and $\mathrm{D} \circ h_{\mathcal{X}} : \Theta \to \mathcal{D}^{\hat{\mathbf{w}}}_\Theta$ are bijective. As a result, we can now identify Θ with $\mathcal{H}_\Theta(\mathcal{X})$ and $\mathcal{D}^{\hat{\mathbf{w}}}_\Theta$. Note also that since $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^{\mathbf{w}}$ is bijective, we can identify $\mathcal{H}(\mathcal{X})$ with $\mathcal{D}^{\mathbf{w}}$.

The fact that Assumption 5.4.2 is sufficient for the identification condition to hold has an important practical implication. Identifiable uniqueness and identification are sometimes equivalent concepts in applications involving well-specified models. For example, when $\mathrm{D}^{\hat{\mathbf{w}}}_0 \in \mathcal{D}^{\mathbf{w}}_\Theta$, Θ is compact and $Q^{\mathcal{D}}_\infty$ is a continuous pre-metric, then identification is both necessary and sufficient for the identifiable uniqueness of $\boldsymbol{\theta}_0$.[15] The results discussed here are thus especially relevant for misspecified models. They are not necessarily interesting otherwise (since identifiable uniqueness would require only verification of 5.4.2).

Indeed, it is precisely when $h_0 \notin \mathcal{H}_\Theta(\mathcal{X}) \Leftrightarrow \mathrm{D}^{\mathbf{w}}_0 \notin \mathcal{D}^{\hat{\mathbf{w}}}_\Theta$ that the present formulation of the problem becomes advantageous. In particular, it is useful to note that given $\mathrm{D}^{\mathbf{x}}_0$, then there exists a functional $Q^{\mathcal{H}}_\infty$ that maps pairs of elements from $\mathcal{H}(\mathcal{X})$ to $\mathbb{R}$, such that $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\mathcal{D}}_\infty(\mathrm{D}^{\mathbf{w}}_0, \mathrm{D}^{\hat{\mathbf{w}}}_{\boldsymbol{\theta}}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q^{\mathcal{H}}_\infty(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$. Writing $Q^{\mathcal{H}}_\infty : \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X}) \to \mathbb{R}^+_0$ is convenient because it conveys the notion of the limiting criterion establishing a divergence $d_{\mathcal{H}}$ on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$. Clearly, $d_{\mathcal{H}}$ is induced by $d_{\mathcal{D}}$ on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ through D according to $d_{\mathcal{H}}(h_1, h_2) = d_{\mathcal{D}}(\mathrm{D}(h_1), \mathrm{D}(h_2)) \,\forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$. Given $h_{\mathcal{X}}$ and $\mathrm{D}^{\mathbf{x}}_0$, the limit $\boldsymbol{\theta}_0$ is thus to be seen as the element in Θ that minimizes the divergence $d_{\mathcal{H}}$ between $h_0 \in \mathcal{H}(\mathcal{X})$ and $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \subseteq \mathcal{H}(\mathcal{X})$. This is stated concisely as $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ where $d_{\mathcal{H}}(h_0, h_{\mathcal{X}}) : \Theta \to \mathbb{R}^+_0$. The employed notion of divergence can be quite general, such as e.g. coinciding with that of a pre-metric, pseudo-metric or quasi-metric. As mentioned before, even though there is no guarantee that $h_0 \in \mathcal{H}_\Theta(\mathcal{X})$, we shall see that under certain conditions $\exists \boldsymbol{\theta}_0 \in \Theta : d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}_0)) < d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta})) \,\forall (\boldsymbol{\theta} \neq \boldsymbol{\theta}_0) \in \Theta$, and hence, that $h_{\mathcal{X}}(\boldsymbol{\theta}_0)$ is the unique best approximation from $\mathcal{H}_\Theta(\mathcal{X})$ to h_0 in $\mathcal{H}(\mathcal{X})$ w.r.t. $d_{\mathcal{H}}$. This implies, under Assumption 5.4.2, that $\boldsymbol{\theta}_0$ is the unique minimizer of $Q^{\mathcal{H}}_\infty$.

Finally, we assume that an appropriate transformation of the limit criterion function yields us with a metric or norm. We emphasize that the only purpose of this assumption is that of retaining the simplicity of the argument, keeping technical requirements to a minimum and allowing us to focus on what is essential. This assumption allows us to make use of the "classical" theorems on existence and uniqueness of best approximations produced in the field of Approximation Theory, which have been naturally obtained for metric, normed and inner product spaces; see Cheney (1982) for a detailed list of existence and uniqueness (and other) accomplish-

[15] The pre-metric is associated here with a divergence that satisfying non-negativity $d_{\mathcal{H}}(h_1, h_2) \geq 0 \,\forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ and identity of indiscernibles $d_{\mathcal{H}}(h_1, h_2) = 0$ if and only if $h_1 = h_2 \,\forall (h_1, h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$.

ments in the field. Even though equivalent results exist for non-metric divergences such as e.g. semi-metrics, pseudo-metrics or quasi-norms, clarity dictates that we consider here only the simpler results available for standard distances.[16] A sufficient requirement in this context is hence that there exists a continuous strictly increasing function g such that $Q_\infty^{\mathcal{H}}$ induces a metric on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$.[17]

Assumption 5.4.3. *There exists a continuous strictly increasing function $g : \mathbb{R} \to \mathbb{R}_0^+$ such that $d_{\mathcal{D}}^* \equiv g \circ Q_\infty^{\mathcal{D}} : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a metric.*

5.5 Strong Unicity of Best Approximations

This section reviews some important results stemming from the field of Approximation Theory. The reader already familiar with this literature might find it preferable to proceed directly to Section 5.6. Observe first the following useful definitions available e.g. in Cheney (1974), Ahuja et al. (1977), Nurberger (1979) and Narang (1981). Let $(\mathcal{B}, d_{\mathcal{B}})$ be a linear metric space. Consider a subset $\mathcal{A} \subset \mathcal{B}$. A *projection mapping* is a set valued map $P_{d_{\mathcal{B}}}^{\mathcal{A}} : \mathcal{B} \to 2^{\mathcal{A}}$ satisfying $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b) := \{a_0 \in \mathcal{A} : d_{\mathcal{B}}(b, a_0) \leq d_{\mathcal{B}}(b, a), a \in \mathcal{A}\}\, \forall b \in \mathcal{B}$, where $2^{\mathcal{A}}$ denotes the power set of $\mathcal{A}$. Note that $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b)$ is the set of elements of best approximation of $b \in \mathcal{B}$ in $\mathcal{A}$, under $d_{\mathcal{B}}$. A set $\mathcal{A} \subset \mathcal{B}$ is then called *proximinal* if $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b)$ is non-empty for every $b \in \mathcal{B}$ and *semi-Chebyshev* if $P_{d_{\mathcal{B}}}^{\mathcal{A}}(b)$ contains at most one element for every $b \in \mathcal{B}$. A set that is both proximinal and semi-Chebyshev is called *Chebyshev*. Note furthermore that a metric space $(\mathcal{B}, d_{\mathcal{B}})$ is said to be *strongly convex* if for every $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ and every $t \in [0, 1]$ there exists a unique $b \in \mathcal{B}$ such that $d_{\mathcal{B}}(b_1, b) = (1 - t)d_{\mathcal{B}}(b_1, b_2)$ and $d_{\mathcal{B}}(b, b_2) = td_{\mathcal{B}}(b_1, b_2)$, i.e. each $t \in [0, 1]$ determines a unique element of the segment $[b_1, b_2] := \{b \in \mathcal{B} : d_{\mathcal{B}}(b_1, b) + d_{\mathcal{B}}(b, b_2) = d(b_1, b_2)\}$. Also, a strongly convex metric space $(\mathcal{B}, d_{\mathcal{B}})$ is said to be *strictly convex* if for every $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ and $r > 0$, $d_{\mathcal{B}}(b_1, b_0) \leq r$, $d_{\mathcal{B}}(b_2, b_0) \leq r$ implies $d_{\mathcal{B}}(b, b_0) < r$ every $b \in]b_1, b_2[:= [b_1, b_2] \backslash \{b_1, b_2\}$ and fixed $b_0 \in \mathcal{B}$.[18]

When a function g exists that satisfies the properties postulated in Assumption 5.4.3, then, the following lemmas adapted from Cheney (1974), Ahuja et al. (1977), Powell (1981, p.4), Narang (1981) and Cheney (1982, p.4), are available to judge on the existence and uniqueness of a best approximation.

Lemma 5.5.1. (Existence on Metric Spaces) *Let $(\mathcal{B}, d_{\mathcal{B}})$ be a metric space and $\mathcal{A} \subseteq \mathcal{B}$ be compact. Then $\mathcal{A}$ is proximinal; i.e. for every $b \in \mathcal{B}$ there exists an element*

[16] These results shed some light on the pathologies identified by Donoho and Liu (1988) concerning the consistency of minimum distance estimators.

[17] As we shall see in section 5.7, it is often straightforward to verify if Assumption 5.4.3 holds.

[18] In a strictly convex metric space $(\mathcal{B}, d_{\mathcal{B}})$ if $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ are two points in the boundary of a sphere, then the open line segment $]b_1, b_2[$ lies strictly inside the sphere.

$a^* \in \mathcal{A}$, a best approximation to b from $\mathcal{A}$, satisfying $d_{\mathcal{B}}(a^*, b) \leq d_{\mathcal{B}}(a, b)\ \forall a \in \mathcal{A}$.

Lemma 5.5.2. (Uniqueness on Metric Spaces) (i) *Let $(\mathcal{B}, d_{\mathcal{B}})$ be a strongly convex metric space and $\mathcal{A} \subseteq \mathcal{B}$ be convex. Then $\mathcal{A}$ is semi-Chebyshev; i.e. there exists at most one element $a^* \in \mathcal{A}$ such that $d_{\mathcal{B}}(a^*, b) \leq d_{\mathcal{B}}(a, b)\ \forall a \in \mathcal{A}$.* (ii) *Let $(\mathcal{B}, d_{\mathcal{B}})$ be a strictly convex metric space. Then $\mathcal{A}$ is semi-Chebyshev.*

The following lemma then follows from combining Lemmas 5.5.1 and 5.5.2 above, and theorem 2 in Ahuja et al. (1977).

Lemma 5.5.3. (Uniqueness on Metric Spaces) (i) *Let $(\mathcal{B}, d_{\mathcal{B}})$ be a strongly convex metric space and $\mathcal{A}$ be a compact convex subset of $\mathcal{B}$. Then $\mathcal{A}$ is Chebyshev; i.e. there exists a unique element $a^* \in \mathcal{A}$ such that $d_{\mathcal{B}}(a^*, b) \leq d_{\mathcal{B}}(a, b)\ \forall a \in \mathcal{A}$.* (ii) *Let $(\mathcal{B}, d_{\mathcal{B}})$ be a strictly convex metric space and $\mathcal{A} \subset \mathcal{B}$ compact. Then $\mathcal{A}$ is Chebyshev.*

Given the linearity of the function spaces considered under the usual definition of addition and multiplication by scalars, it is often beneficial to work on normed vector spaces. Some estimators might have limiting criterion functions $Q_\infty^{\mathcal{H}}$ for which $g \circ Q_\infty^{\mathcal{H}}$ is a metric on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ but not a norm (since the latter requires also homogeneity and translation invariance). When $g \circ Q_\infty^{\mathcal{H}}$ is a norm on $\mathcal{H}(\mathcal{X})$ however, simpler results from Approximation Theory are available for the uniqueness of best approximations. For this reason the following assumption is also introduced.

Assumption 5.5.1. *There exists a continuous strictly increasing function $g : \mathbb{R} \to \mathbb{R}_0^+$ such that $g \circ Q_\infty^{\mathcal{D}}(\mathrm{D}, \mathrm{D}') \equiv \| \mathrm{D} - \mathrm{D}' \|_{\mathcal{D}}\ \forall (\mathrm{D}, \mathrm{D}') \in \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$ where $\| \cdot \|_{\mathcal{D}} \colon \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a norm.*

Consider now the natural extensions of the definition of strictly convex metric space to normed vector spaces. Let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a normed vector space. Then $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ is said to be strictly convex if for every $(b_1, b_2) \in \mathcal{B} \times \mathcal{B}$ satisfying $\| b_1 \|_{\mathcal{B}} = \| b_2 \|_{\mathcal{B}} = 1$ the inequality $\| (1-t)b_1 + tb_2 \|_{\mathcal{B}} < 1$ holds for every $t \in]0, 1[$.

The following lemmas, which follow from those above for metric spaces, are adapted from Powell (1981, p.6,13-15) and Cheney (1982, p.20,23), and establish a few useful results on the existence and uniqueness of best approximations.

Lemma 5.5.4. (Existence on Normed Spaces) *Let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a normed space and $\mathcal{A}$ a finite-dimensional subset of $\mathcal{B}$. Then $\mathcal{A}$ is proximinal.*

Lemma 5.5.5. (Uniqueness on Normed Spaces) (i) *Let $\mathcal{A} \subset B$ be a compact and strictly convex set in a normed linear space $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$. Then $\mathcal{A}$ is Chebyshev.* (ii) *Let $\mathcal{A} \subset B$ be a convex set in a strictly convex normed linear space $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$. Then $\mathcal{A}$ is semi-Chebyshev.* (iii) *Let $\mathcal{A} \subset B$ be a finite dimensional subspace of $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$. Then $\mathcal{A}$ is Chebyshev.*

Here it is important to point out that e.g. the well known L^1 and sup norms do not satisfy the strict convexity property of Lemma 5.5.5 (nor that of Lemma 5.5.2 in the induced metrics). Fortunately, the well known Haar condition allows us to overcome this limitation.

Definition 5.5.1. (Haar Condition) *A system of functions* $\{\psi_1, ..., \psi_n\}$ *with* $\psi_i : \mathcal{A} \subset \mathbb{R} \to \mathbb{R}$, $i = 1, ..., n$ *is said to satisfy the Haar condition on* $\mathcal{A}$ *if each* $\psi_i \in C(\mathcal{A})$, *the space of continuos functions on* $\mathcal{A}$, *for* $i = 1, ..., n$, *and if every set of* n *vectors of the form* $[\psi_1(a), ..., \psi_n(a)]$, $a \in \mathcal{A}$ *is independent; i.e. if for any given collection* $(a_1, ..., a_n) \in \times_{i=1}^{n} \mathcal{A}$, $a_i \neq a_j \, \forall i \neq j, i = 1, ...n, j = 1, ...n$, *the system has non-vanishing Vandermonde's determinant.*

A subspace $\mathcal{H}_\Theta(\mathcal{X}) \subset C(\mathcal{X})$ of generalized polynomials spanned by a system of functions $\{\psi_1, ..., \psi_n\}$ satisfying the Haar condition is called a *Haar subspace* of $C(\mathcal{X})$. The following lemma is adapted from Cheney (1982, p.81,219) and Powell (1981, 80,170). It is suitable for both L^1 and sup norm approximations.

Lemma 5.5.6. (Haar's Unicity theorem) *Let* $\mathcal{H}_\Theta(\mathcal{X})$ *be a Haar subspace of the spaces* $(C(\mathcal{X}), \| \cdot \|_1)$ *or* $(C(\mathcal{X}), \| \cdot \|_\infty)$ *and* $\mathcal{X}$ *a compact Hausdorff space. Then,* $\mathcal{H}_\Theta(\mathcal{X})$ *is Chebyshev.*

The Haar condition offers more than just a unicity characterization of best approximations on normed linear subspaces of $C(\mathcal{X})$. Under certain conditions, the element of best approximation from a Haar subspace is characterized by the strong unicity property. This property is relevant in the present context since the identifiable uniqueness condition in Assumption 5.3.3 can be derived from it. Following Newman and Shapiro (1963) and Cheney (1982, p.80), let $(\mathcal{B}, \| \cdot \|_\mathcal{B})$ be a normed linear space and $a \in \mathcal{A} \subseteq \mathcal{B}$ an element of best approximation to $b_0 \in \mathcal{B}$ from $\mathcal{A}$. Then, a is said to be *strongly unique* if $\exists \, \gamma(b_0) > 0 : \| \, b_0 - a' \, \| > \| \, b_0 - a \, \| + \gamma \, \| \, a - a' \, \| \;\; \forall a' \in \mathcal{A}$.

Lemma 5.5.7. (Strong Unicity in Normed Linear Spaces) *Let* $\mathcal{H}_\Theta(\mathcal{X})$ *be a Haar subspace of* $(C(\mathcal{X}), \| \cdot \|_\infty)$ *and* $\mathcal{X}$ *a compact Hausdorff space. Then, for every* $h_0 \in C(\mathcal{X})$ *the element* $h \in \mathcal{H}_\Theta(\mathcal{X})$ *of best approximation to* $h_0 \in C(\mathcal{X})$ *is strongly unique; i.e. there exists a generalized polynomial* $h \in \mathcal{H}_\Theta(\mathcal{X})$, $h = \sum_{i=1}^{n} \theta_i \psi_i$ *where* $\{\psi_1, ..., \psi_n\}$ *satisfy the Haar condition, such that there exists* $\gamma(h_0) > 0 : \| \, h_0 - h' \, \|_\infty > \| \, h_0 - h \, \|_\infty + \gamma \, \| \, h - h' \, \|_\infty \;\; \forall h' \in \mathcal{H}_\Theta(\mathcal{X})$.

Unfortunately, Lemma 5.5.7 is available only under the sup norm. Furthermore, it is known since Wulbert (1971) that strong unicity of elements of best approximation is generally not available in smooth Banach spaces.

This holds in particular in $L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_\mathcal{E})$ spaces, with $1 < p < \infty$, where $(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_\mathcal{E})$ is a given measure space. Fortunately, the identifiable uniqueness

property of Assumption 5.3.3 can also be derived from the concept of *strong unicity of order* α. Following Angelos and Egger (1984) and Lin (1989), let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a Banach space and $a \in \mathcal{A} \subseteq \mathcal{B}$ be an element of best approximation to $b_0 \in \mathcal{B}$ from $\mathcal{A}$. Then, a is said to be *strongly unique of order* α ($\alpha > 1$) if $\exists\, \gamma(b_0) > 0 : \| b_0 - a' \| > \| b_0 - a \| + \gamma \| a - a' \|^{\alpha} \ \forall a' \in \mathcal{A}$.

The following lemma, adapted from Angelos and Egger (1984) and Lin (1989), reveals that this strong unicity property holds for finite-dimensional subspaces of $L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$ smooth Banach spaces ($1 < p < \infty$), or general subspaces of uniformly convex Banach spaces of type p. Note that a Banach space $(\mathcal{A}, \| \cdot \|)$ is said to be *uniformly convex* (Clarkson (1936)) if for every $0 < \epsilon \leq 2$ there exists a $\delta(\epsilon) > 0$ such that having $\| a_1 \| = \| a_2 \| = 1$ and $\| a_1 - a_2 \| \geq \epsilon$ implies $\| (a_1 + a_2)/2 \| \leq 1 - \delta(\epsilon)$.

The function $\delta(\epsilon) : (0, 2] \to [0, 1]$ defined as $\delta(\epsilon) = \inf\{ 1 - 1/2 \| a_1 + a_2 \| \mid \| a_1 \| \leq 1, \| a_2 \| \leq 1, \| a_1 - a_2 \geq \epsilon \}$ is called the *modulus of convexity* of the Banach space $(\mathcal{A}, \| \cdot \|)$, and this space is said to be *uniformly convex of power type* p if there exists $\Delta > 0$ such that $\delta(\epsilon) \geq \Delta \epsilon^p$.

The following lemma uses also a result of Hanner (1956) showing that L^p spaces with $1 < p < \infty$ are uniformly convex of power type $\max\{2, p\}$, and the fact that strictly convex normed linear spaces are also uniformly convex; see e.g. Cheney (1974) or Cheney (1982, p.23).

Lemma 5.5.8. (Strong Unicity of Order α in Normed Linear Spaces) *(i) Let $\mathcal{A}$ be a finite-dimensional subspace of an $L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$ space with $1 < p < \infty$ and $\mathcal{E} \subseteq \mathbb{R}^{n_{\mathcal{E}}}$. Then, the element $a \in \mathcal{A}$ of best approximation to $b \in L^p(\mathcal{E}, \mathfrak{B}(\mathcal{E}), \mu_{\mathcal{E}})$, when it exists, is strongly unique of order $\alpha = \max\{p, 2\}$. (ii) Let $(\mathcal{B}, \| \cdot \|_{\mathcal{B}})$ be a uniformly convex Banach space of power type p and let $\mathcal{A}$ be a subspace of $\mathcal{B}$. Then, an element $a \in \mathcal{A}$ of best approximation to $b \in \mathcal{B}$, when it exists, is strongly unique of order p.*

Also, similar results to those obtained above are available under weaker conditions on the employed notion of distance. Examples include Romaguera and Sanchis (2000) that deal with quasimetric spaces and S. and C. (2006) that work with asymmetric normed linear spaces.

While these formulations might offer more generality, we manage to achieve a significant simplification by restricting ourselves to the former case where $Q_{\infty}^{\mathcal{H}}$ induces a metric or norm on $\mathcal{H}(\mathcal{X})$. The reader should nevertheless bear in mind the limitations introduced by the simplifying assumption just mentioned. This is important as this restriction might prove to be relevant in several applications.

5.6 Consistency Restated

Finally, we are ready to restate the consistency results of Section 5.3 using alternative conditions. We note in particular that Assumption 5.3.3 (identifiable uniqueness of $\boldsymbol{\theta}_0$) and Assumption 5.3.4 (uniqueness of $\boldsymbol{\theta}_0$) used in Lemmas 5.3.2 and 5.3.3 to obtain the consistency of $\hat{\boldsymbol{\theta}}_T$ can now be substituted by sets of sufficient conditions that make use of the problem formulation discussed in Section 5.4 and the lemmas of Section 5.5 on the unicity of best approximations. Under the more restrictive assumptions of Lemma 5.3.3, which impose the compactness of Θ and continuity of Q_∞, showing the uniqueness of $\boldsymbol{\theta}_0$ is enough to obtain the consistency of $\hat{\boldsymbol{\theta}}_T$ since in this setting, a unique $\boldsymbol{\theta}_0$ is automatically identifiably unique. In this simpler case, we will need only to make use of those lemmas establishing the uniqueness of best approximations covered in Section 5.5. It is under the less restrictive conditions of Lemma 5.3.2 that the results on strong unicity of best approximations become important since, in that case, $Q_\infty(\boldsymbol{\theta}_0)$ must be shown to be well separated without the aid of the compactness of Θ or the continuity of Q_∞.

As we have seen in the previous section, the uniqueness of $\boldsymbol{\theta}_0$ can be established either in the context of metric spaces or that of normed linear spaces. Depending on the problem, each formulation will be more or less advantageous in terms of verification.[19] Assumptions 5.6.1 and 5.6.2 below establish the conditions from which the uniqueness of $\boldsymbol{\theta}_0$ will be derived. These make use of the fact that every convex proximinal set is Chebyshev and are stated for future reference. Assumptions 5.6.3, 5.6.4 and 5.6.5 establish useful conditions for directly deriving the identifiable uniqueness of $\boldsymbol{\theta}_0$.

Assumption 5.6.1. (i) $(\mathcal{H}(\mathcal{X}), d^*_{\mathcal{H}})$ *is a strongly convex metric space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a compact convex subset of* $\mathcal{H}(\mathcal{X})$*; or* (ii) $(\mathcal{H}(\mathcal{X}), d^*_{\mathcal{H}})$ *is a strictly convex metric space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a compact subset of* $\mathcal{H}(\mathcal{X})$*.*

Assumption 5.6.2. (i) $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}})$ *is a normed linear space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a compact strictly convex subset of* $\mathcal{H}(\mathcal{X})$*; or* (ii) $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}})$ *is a strictly convex normed vector space and* $\mathcal{H}_\Theta(\mathcal{X})$ *a finite dimensional convex subset of* $\mathcal{H}(\mathcal{X})$*.*

Assumption 5.6.3. $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}}) = (C(\mathcal{X}), \|\cdot\|_\infty)$ *where* $\|\cdot\|_\infty$ *denotes the supremum norm, and for every* $\boldsymbol{\theta} \in \Theta$*, the elements* $h(\cdot;\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ *accept a generalized polynomial representation* $h(\cdot,\boldsymbol{\theta}) = \sum_{i=1}^{n_h} \boldsymbol{\theta}_i h_i$ *where* $\{h_1, ..., h_n\}$ *satisfies the Haar condition.*

[19] In particular, while simpler results are available for norms, the limiting criterion $Q^{\mathcal{D}}_\infty$ that induces a metric on $\mathcal{D}^{\mathbf{w}}$ must also be homogeneous and translation invariant to establish a norm on the vector space.

Assumption 5.6.4. $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}}) = L^p(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mu_{\mathcal{X}})$ *with* $1 < p < \infty$*, so that* $\|\cdot\|_{\mathcal{H}}$ *satisfies* $\| h \|_{\mathcal{H}} = \left(\int_{\mathcal{X}} |h|^p d\mu \right)^{1/p} \forall h \in \mathcal{H}(\mathcal{X})$ *with* $1 < p < \infty$*. Furthermore,* $\mathcal{H}_\Theta(\mathcal{X})$ *is a finite dimensional subspace of* $\mathcal{H}(\mathcal{X})$*.*

Assumption 5.6.5. $(\mathcal{H}(\mathcal{X}), \|\cdot\|_{\mathcal{H}})$ *is a uniformly convex Banach space of power type* p *and* $\mathcal{H}_\Theta(\mathcal{X})$ *is a closed convex subspace of* $\mathcal{H}(\mathcal{X})$*.*

Finally, we derive the uniqueness of $\boldsymbol{\theta}_0$ from the properties of the limiting criterion function $Q_\infty^{\mathcal{H}}$ and the space of parametric functions $\mathcal{H}_\Theta(\mathcal{X})$ implied by both the parameterization mapping $h_{\mathcal{X}}$ and the parameter space Θ. Theorem 5.6.1 below addresses uniqueness in the context of metric-inducing limiting criteria $Q_\infty^{\mathcal{H}}$.

Theorem 5.6.1. (Uniqueness for Metric Limit Criteria) *Let Assumptions 5.4.1, 5.4.2, 5.4.3 and 5.6.1 hold. Then* $Q_\infty : \Theta \to \mathbb{R}$ *has a unique minimum at* $\boldsymbol{\theta}_0$*.*

Proof. See Section 5.9 □

Now, in light of Lemma 5.3.3, the a.s. convergence of $\hat{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}_0$ follows immediately as corollary under the added influence of Assumptions 5.3.1 and 5.3.2.

Corollary 5.6.1. *Let Assumptions 5.3.1, 5.3.2, 5.4.1, 5.4.2, 5.4.3 and 5.6.1 hold. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta})$*. Then* $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $T \to \infty$*.*[20]

Accordingly, Theorem 5.6.2 below addresses the uniqueness of $\boldsymbol{\theta}_0$ in the context of norm-inducing limiting criteria $Q_\infty^{\mathcal{H}}$.

Theorem 5.6.2. (Uniqueness for Norm Limit Criteria) *Let Assumptions 5.4.1, 5.4.2, 5.5.1 and 5.6.2 hold. Then* $Q_\infty : \Theta \to \mathbb{R}$ *has a unique minimum at* $\boldsymbol{\theta}_0$*.*

Proof. See Section 5.9 □

Again, $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ follows immediately as a corollary when Assumptions 5.3.1 and 5.3.2 also hold.

Corollary 5.6.2. *Let Assumptions 5.3.1, 5.3.2, 5.4.1, 5.4.2, 5.5.1 and 5.6.2 hold. Define* $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ *such that* $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta} \in \Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta})$*. Then* $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ *as* $T \to \infty$*.*

[20] It is well known that standard consistency proofs apply also to approximate extremum estimators, thus eliminating the need to impose the existence conditions postulated in Assumption 5.3.1 and substituting it by more general conditions for the existence of measurable approximate minimizers of the criterion function of interest (see e.g. Brown and Purves (1973)).

When the assumptions of Lemma 5.3.3 are too restrictive, it is possible to work with those of Lemma 5.3.2 instead by verifying that identifiable uniqueness follows essentially from the stricter conditions of Assumptions 5.4.2 and 5.5.1, plus either 5.6.3, 5.6.4 or 5.6.5. In particular, it is possible to relax the assumptions of compactness of Θ and continuity of $Q_\infty : \Theta \to \mathbb{R}$. This however, is not to be done without the further qualification stated in Assumption 5.6.6 below.

Assumption 5.6.6. *$h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is an open map.*[21]

The following theorem establishes the relation between the concepts of strong unicity found in the previous section and that of identifiable uniqueness used in Lemma 5.3.2.

Theorem 5.6.3. (Strong Unicity Implies Identifiable Uniqueness) *Let Assumptions 5.4.1, 5.4.2, 5.5.1 and 5.6.6 be satisfied. Then $Q_\infty : \Theta \to \mathbb{R}$ has an identifiably unique minimizer $\boldsymbol{\theta}_0$ if either Assumption 5.6.3, 5.6.4 or 5.6.5 hold.*

Proof. See Section 5.9. □

This time $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ follows as corollary of Theorem 5.6.3 and Lemma 5.3.2.

Corollary 5.6.3. *Let Assumptions 5.3.2, 5.4.1, 5.4.2, 5.5.1 and 5.6.6 be satisfied. Define $\hat{\boldsymbol{\theta}}_T : \Omega \to \Theta$ such that $\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_\mathbf{T}, \mathbf{x}_\mathbf{T}; \boldsymbol{\theta})$. Then $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $T \to \infty$ if either Assumption 5.6.3, 5.6.4 or 5.6.5 hold.*

Finally, we use a number of simple examples that illustrate how to verify that the conditions for uniqueness and identifiable uniqueness postulated in Assumptions 5.3.3 and 5.3.4 hold.

5.7 Some Examples

In the previous sections of this chapter we obtained the desired results essentially by decomposing the mapping of elements from Θ to $\mathbb{R}$, compounded in the limiting objective function $Q_\infty : \Theta \to \mathbb{R}$, into three sub-mappings that are easier to handle. We thus obtained a more transparent account of the structure of the extremum estimation problem in nonlinear regression models. The three sub-maps are: *(i)* the so-called *parameterization mapping* $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$, *(ii)* the *probability measure map* $\mathrm{D} : \mathcal{H}(\mathcal{X}) \to \mathcal{D}^\mathbf{w}$, and finally, *(iii)* the divergence criterion function $Q_\infty^\mathcal{D} : \mathcal{D}^\mathbf{w} \times \mathcal{D}^\mathbf{w} \to \mathbb{R}_0^+$.

[21] A sufficient condition for the openess of $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ is that its inverse $h_\mathcal{X}^{-1} : \mathcal{H}_\Theta(\mathcal{X}) \to \Theta$ be continuous in $h \in \mathcal{H}_\Theta(\mathcal{X})$. Also, note that *(i)* the existence of the inverse function $h_\mathcal{X}^{-1}$ is assured by the bijectiveness of $h_\mathcal{X}$, and that *(ii)* in the special case where $h_\mathcal{X}$ is also continuous, then $h_\mathcal{X}$ is an homeomorphism.

Simple conditions on each of these sub-maps, as well as the sets Θ, $\mathcal{H}_\Theta(\mathcal{X})$ and $\mathcal{D}_\Theta^{\hat{\mathbf{w}}}$, were shown to ensure the identifiable uniqueness of $\boldsymbol{\theta}_0$. We now review very briefly simple examples of regression models and extremum estimators satisfying the above mentioned properties. The purpose of this section is only that of clarifying the nature of the Assumptions 5.4.2 to 5.6.6. To remain short and concise, we discuss only a few cases for which verification is straightforward. The interesting cases are likely to be those requiring a more intricate argument. These however are left to be found by researchers having specific applications in mind.

5.7.1 Parameterization Mapping: Some Regression Models

Several immediate examples of regression models can be devised for which the injective and open properties of the parameterization mapping $h_\mathcal{X}$ hold (Assumptions 5.4.2 and 5.6.6) and where properties such as compactness, convexity, closedness, finite dimensionality and Haar characterization of $\mathcal{H}_\Theta(\mathcal{X})$ (in Assumptions 5.6.1, 5.6.2, 5.6.3, 5.6.4 or 5.6.5) are trivially satisfied. As we shall see, it is generally easy to derive the properties of $\mathcal{H}_\Theta(\mathcal{X})$ from those of Θ, whose qualities are defined by the researcher in any given application.

Note first that *the bijective nature of* $h_\mathcal{X}$ (implied by Assumption 5.4.2) is generally easily derived in this simple regression framework. This is true for instance in models involving polynomial, exponential, logarithmic, trignometric or power functions, that satisfy simple regularity conditions. Note for example that for regression functions that are analytic on the domain of interest, i.e. $h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \equiv C_\Theta^\omega(\mathcal{X})$, the bijective nature of $h_\mathcal{X}$ follows immediately from the fact that each element of $C_\Theta^\omega(\mathcal{X})$ has a power-series representation. The uniqueness of this representation, and hence the bijective nature of $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$, then follows immediately from the uniqueness of power series.[22]

The finite dimensionality of $\mathcal{H}_\Theta(\mathcal{X})$ (stated in Assumptions 5.6.2 and 5.6.4) is implied by the finite dimensionality of Θ (which holds in several applications) given the identification of $\mathcal{H}_\Theta(\mathcal{X})$ with Θ (a consequence of $h_\mathcal{X}$ being bijective). This is true e.g. for the case of polynomial regressions $h_\mathcal{X}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \equiv \mathcal{P}_\Theta^k$, $k \in \mathbb{N}$.

The compactness of $\mathcal{H}_\Theta(\mathcal{X})$ (Assumptions 5.6.1 and 5.6.2) is easily obtained, for instance, under the continuity of $h_\mathcal{X}$ and the compactness of Θ. Here note that, for example, given a regression model of the form $h(x_t; \theta_1, \theta_2, \theta_3) = \theta_1 + \theta_2 \exp(-\theta_3 x_t)$, the continuity of $h_\mathcal{X} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ holds for a large class of metric or norm functions with which Θ and $\mathcal{H}_\Theta(\mathcal{X})$ are possibly equipped, and it is immediately satisfied for

[22] In multi-index notation (see e.g. Krantz and Parks (1992, p.25)), let $h(\mathbf{x}_t; \boldsymbol{\theta}) = \sum_{|\boldsymbol{\alpha}|=0}^\infty \boldsymbol{\theta}_{\boldsymbol{\alpha}} \mathbf{x}_t^{\boldsymbol{\alpha}}$ $\forall\, \mathbf{x}_t \in \mathcal{X}$ and $h(\mathbf{x}_t; \boldsymbol{\theta}') = \sum_{|\boldsymbol{\alpha}|=0}^\infty \boldsymbol{\theta}'_{\boldsymbol{\alpha}} \mathbf{x}_t^{\boldsymbol{\alpha}}$ $\forall\, \mathbf{x}_t \in \mathcal{X}$. Then, $h(\mathbf{x}_t; \boldsymbol{\theta}) = h(\mathbf{x}_t; \boldsymbol{\theta}')$ $\forall\, \mathbf{x}_t \in \mathcal{X}$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

polynomial regression functions $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{P}_{\Theta}^{k}$ regardless of the metric or norm defined on these spaces.

The convexity of $\mathcal{H}_{\Theta}(\mathcal{X})$ (used in Assumption 5.6.1, 5.6.2 and 5.6.5) can be easily obtained from the convexity of Θ for a large class of parameterization mappings. For example, in the case of a polynomial regression of order k, when $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_{\Theta}(\mathcal{X}) \equiv \mathcal{P}_{\Theta}^{k}$, we have that, for every $(h_{\mathcal{X}}(\boldsymbol{\theta}_1), h_{\mathcal{X}}(\boldsymbol{\theta}_2)) \in \mathcal{H}_{\Theta}(\mathcal{X}) \times \mathcal{H}_{\Theta}(\mathcal{X})$ and every $\tau \in [0,1]$, the function $(\tau h_{\mathcal{X}}(\boldsymbol{\theta}_1) + (1-\tau) h_{\mathcal{X}}(\boldsymbol{\theta}_2))$ belongs to $\mathcal{H}_{\Theta}(\mathcal{X})$ and takes the form $h_{\mathcal{X}}(\boldsymbol{\theta}_3)$ with $\boldsymbol{\theta}_3 = \tau\boldsymbol{\theta}_1 + (1-\tau)\boldsymbol{\theta}_2$.

The closedness of $\mathcal{H}_{\Theta}(\mathcal{X})$ (used in Assumption 5.6.5) can be easily obtained, for instance, under the closedness of Θ and the continuity of $h_{\mathcal{X}}^{-1} : \mathcal{H}(\mathcal{X}) \to \Theta$.[23] The continuity of the inverse parameterization mapping $h_{\mathcal{X}}^{-1}$ is easily obtained for a large class of regression models. It holds, for example, on regressions models based on power functions $h(x_t; \theta_1, \theta_2) = \theta_1 x_t^{\theta_2}$ under a large class of norms on Θ and $\mathcal{H}_{\Theta}(\mathcal{X})$. Once more, it also holds for polynomial regressions under arbitrary norms.

The openess of $h_{\mathcal{X}} : \Theta \to \mathcal{H}_{\Theta}(\mathcal{X})$ *(postulated in Assumption 5.6.6)* is also implied by the continuity of the inverse map $h_{\mathcal{X}}^{-1}$. Hence, the previous argument holds as well.[24]

The Haar characterization of $\mathcal{H}_{\Theta}(\mathcal{X})$ (Assumption 5.6.3) has been obtained for large classes of functions. For example, $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{P}_{\Theta}^{k}(\mathcal{X})$ satisfies trivially the Haar condition.[25]

5.7.2 Limit Divergence Criterion: Illustrative Estimators

We now observe how the properties of the divergence map $Q_{\infty}^{\mathcal{H}} : \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X}) \to \mathbb{R}$ implicitly defined in Assumptions 5.4.3, 5.5.1 and 5.6.1-5.6.5 are directly obtained from those of the estimation procedure employed. In particular, we discuss the verification of the simplifying assumption that $g \circ Q_{\infty}^{\mathcal{H}}$ (g strictly increasing) be a metric or norm, and that it be either, strongly convex, strictly convex, uniformly convex, of the L^p type ($p < \infty$), or the supremum norm.

The existence of a metric/norm $g \circ Q_{\infty}^{\mathcal{H}}$ *on* $\mathcal{H}(\mathcal{X})$ (established in Assumptions 5.4.3 and 5.5.1 and used in Assumptions 5.6.1-5.6.5) is immediate for the class of minimum distance estimators (e.g. the minimum Hellinger distance estimator),

[23] Existence of $h_{\mathcal{X}}^{-1}$ is assured by the bijective nature of $h_{\mathcal{X}}$. A bijective map is closed if and only if it is open. The inverse of a continuous map is open.

[24] An obvious sufficient condition is that $h_{\mathcal{X}}$ be a homeomorphism, i.e. that $h_{\mathcal{X}}$ be bijective, continuous with continuous $h_{\mathcal{X}}^{-1}$. Note that the homeomorphic nature of $h_{\mathcal{X}}$ can be obtained by letting (Θ, d_{Θ}^{*}) be a metric space with d_{Θ}^{*} induced by $h_{\mathcal{X}}^{-1}$ so that $h_{\mathcal{X}}$ is automatically isometric and also an isometric isomorphism.

[25] Power monomials satisfy the Haar condition. The system $\{1, \mathbf{x}_t, ..., \mathbf{x}_t^{\mathbf{k}}\}$ has non-vanishing Vandermonde's determinant $VD[a_1, ..., a_k] \neq 0$ $(a_1, ..., a_k) \in \times_{i=1}^{k} \mathbb{R}_i^{n_x}$ and hence it satisfies the Haar condition.

since by definition, these estimators are such that $Q_\infty^{\mathcal{D}}$ takes the form of a distance on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$. As observed in Section 5.4, a metric or norm is then induced on $\mathcal{H}(\mathcal{X})$ by the bijective mapping D. For many other estimators, in particular those that are not directly obtained as distance minimizers, it is often easy to find a strictly increasing function g such that $g \circ Q_\infty^{\mathcal{H}}$ defines a distance on $\mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$. For example, it is well known that under appropriate regularity conditions, the least squares estimator,

$$\hat{\boldsymbol{\theta}}_T^{LS} := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(y_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}; \boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{T}\sum_{t=1}^{T} e_t^2$$

with $\sum_{t=1}^T e_t^2 \equiv \sum_{t=1}^T \left[y_t - \hat{y}_t\right]^2 \equiv \sum_{t=1}^T \left[h_0(\mathbf{x}_t) - h(\mathbf{x}_t;\boldsymbol{\theta})\right]^2$ is such that,

$$\boldsymbol{\theta}_0^{LS} := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) := \arg\min_{\boldsymbol{\theta}\in\Theta} \int_{\mathcal{X}} \mathrm{D}_0^{\mathbf{x}}(\mathbf{x}_t)\left[h_0(\mathbf{x}_t) - h(\mathbf{x}_t;\boldsymbol{\theta})\right]^2 d\mathbf{x}_t,$$

see e.g. White (1980b). This implies that $\boldsymbol{\theta}_0^{LS} = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ where $d_{\mathcal{H}}$ is a divergence.[26] Immediately, taking $g \circ d_{\mathcal{H}}(h_1,h_2) = \sqrt{d_{\mathcal{H}}(h_1,h_2)} \equiv \| h_1 - h_2 \|_{\mathcal{H}}$ for every $(h_1,h_2) \in \mathcal{H}(\mathcal{X})$ implies that $\| \cdot \|_{\mathcal{H}}$ is the well known L^2 norm where $\| h(\mathbf{x}) \|_{\mathcal{H}} = \left(\int_{\mathcal{X}} |h|^2 d_{\mathbf{x}}\right)^{1/2}$. Hence, Assumption 8 holds (and 7 as well by the induced metric) and $\boldsymbol{\theta}_0$ can be described as minimizer of $\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}$ on $(\mathcal{H}(\mathcal{X}), \| \cdot \|_{\mathcal{H}}) \equiv (\mathcal{H}(\mathcal{X}), L^2)$, i.e.,

$$\boldsymbol{\theta}_0^{LS} = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left(\int_{\mathcal{X}} \mathrm{D}_0^{\mathbf{x}}(\mathbf{x}_t)\left[h_0(\mathbf{x}_t) - h(\mathbf{x}_t;\boldsymbol{\theta})\right]^2 d\mathbf{x}_t \right)^{1/2}.$$

The strict convexity of $d_{\mathcal{H}}^*$ or $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$ on $\mathcal{H}_\Theta(\mathcal{X})$ (Assumption 5.6.2) is generally easy to verify and it holds e.g. for the minimum Hellinger distance and least squares estimators just mentioned above (see e.g. Donoho and Liu (1988) and Powell (1981) respectively). Note also that in this case the *strong convexity of $d_{\mathcal{H}}^* \equiv g \circ Q_\infty^{\mathcal{H}}$* (used in Assumption 5.6.1) is immediately obtained since the later is by construction implied by the former (see Cheney (1974), Ahuja et al. (1977) and Narang (1981)). This is also true of *uniform convexity of power type p of $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$* (Assumption 5.6.5) and *L^p representation of $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$* (Assumption 5.6.4) in the case of least squares estimation (see Cheney (1974) or Cheney (1982, p.23)).

The supremum representation of $\| \cdot \|_{\mathcal{H}} \equiv g \circ Q_\infty^{\mathcal{H}}$ (Assumption 5.6.3) is considerably more restrictive (as mentioned before) and holds for minimax estimators.

[26] The least squares divergence, $d_{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ satisfies non-negativity $d_{\mathcal{H}}(h_1,h_2) \geq 0 \; \forall (h_1,h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ and identity of indiscernibles $d_{\mathcal{H}}(h_1,h_2) = 0$ if and only if $h_1 = h_2 \; \forall (h_1,h_2) \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$, but not symmetry or sub-additivity.

5.8 Final Remarks

In this chapter we have illustrated the possibility of using results from Approximation Theory to verify the assumption of identifiable uniqueness commonly used to obtain consistency of extremum estimators. We made use only of simple intuitive results on the (strong) uniqueness of best approximations. Clearly, much more can be done in extending these results to a larger class of extremum estimators and regression models. Here, generality was sacrificed in favor of conciseness and simplicity, but it should be kept in mind that, in this context, we could be as general as Approximation Theory allows us to be. In particular, these results extend immediately to various models outside the regression framework and the notion of distance function can be easily weakened to include non-metric divergences.

5.9 Proofs

5.9.1 Theorem 5.6.1

Proof. Assumption 5.4.1 implies that,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$$

and according to Assumption 5.4.3,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{D}}^*(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$$

where $d_{\mathcal{D}}^* : \mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a metric defined on $\mathcal{D}^{\mathbf{w}} \times \mathcal{D}^{\mathbf{w}}$ as $d_{\mathcal{D}}^* \equiv g \circ Q_{\mathcal{D}}$ with $g : \mathbb{R} \to \mathbb{R}_0^+$ a strictly increasing function. Now given Assumption 5.4.2, we have $d_{\mathcal{D}}^*(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}) \equiv d_{\mathcal{H}}^*(h_0, h(\cdot, \boldsymbol{\theta}))\ \forall\, \boldsymbol{\theta} \in \Theta$ by construction since $d_{\mathcal{H}}^* : \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X}) \to \mathbb{R}_0^+$ is a metric defined on $\mathcal{H}(\mathcal{X})$ according to $d_{\mathcal{H}}^*(h, h') \equiv d_{\mathcal{D}}^*(\mathrm{D}(h), \mathrm{D}(h'))\ \forall\, (h, h') \in \mathcal{H}(\mathcal{X}) \times \mathcal{H}(\mathcal{X})$ and hence $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}^*(h_0, h(\cdot, \boldsymbol{\theta}))$ holds true. Finally, according to Lemmas 5.5.1, 5.5.2 and 5.5.3, Assumption 5.6.1 implies that for every $h_0 \in \mathcal{H}(\mathcal{X})$, there exists a unique $h \in \mathcal{H}_\Theta(\mathcal{X})$ satisfying,

$$d_{\mathcal{H}}^*(h_0, h) \leq d_{\mathcal{H}}^*(h_0, h')\ \forall\, h' \in \mathcal{H}_\Theta(\mathcal{H}).$$

Given the bijective nature of the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ postulated in Assumption 5.4.2, it follows that there exists a unique $\boldsymbol{\theta} \in \Theta$ satisfying,

$$d_{\mathcal{H}}^*(h_0, h_{\mathcal{X}}(\boldsymbol{\theta})) \leq d_{\mathcal{H}}^*(h_0, h(\cdot; \boldsymbol{\theta}'))\ \forall\, \boldsymbol{\theta}' \in \Theta;$$

i.e. $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} d_{\mathcal{H}}^*(h_0, h(\cdot, \boldsymbol{\theta})) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ is unique. □

5.9.2 Theorem 5.6.2

Proof. The argument follows essentially that of Theorem 5.6.1. Assumption 5.4.1 ensures that attention is restricted to the class of extremum estimators satisfying $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty(\boldsymbol{\theta}) \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ and by Assumption 5.5.1,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| \mathrm{D}_0^{\mathbf{w}} - \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \|_{\mathcal{D}}$$

where $\| \cdot \|_{\mathcal{D}}$: $\mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a norm defined on $\mathcal{D}^{\mathbf{w}}$ as $\| \cdot \|_{\mathcal{D}} \equiv g \circ Q_\infty^{\mathcal{D}}$ with $g : \mathbb{R} \to \mathbb{R}_0^+$ a strictly increasing transformation. Now given Assumption 5.4.2, we have $\| \mathrm{D}_0^{\mathbf{w}} - \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \|_{\mathcal{D}} \equiv \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}} \ \forall \boldsymbol{\theta} \in \Theta$ by construction, since $\| \cdot \|_{\mathcal{H}}$: $\mathcal{H}(\mathcal{X}) \to \mathbb{R}_0^+$ is a norm defined on $\mathcal{H}(\mathcal{X})$ according to $\| h \|_{\mathcal{H}} = \| \mathrm{D}(h) \|_{\mathcal{D}} \ \forall h \in \mathcal{H}(\mathcal{X})$, and hence $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$ holds true. Finally, according to Lemmas 5.5.4 and 5.5.5, Assumption 5.6.2 implies that for every $h_0 \in \mathcal{H}(\mathcal{X})$, there exists a unique $h \in \mathcal{H}_\Theta(\mathcal{X})$ satisfying,

$$\| h_0 - h \|_{\mathcal{H}} \leq \| h_0 - h' \|_{\mathcal{H}} \ \forall h' \in \mathcal{H}_\Theta(\mathcal{H}).$$

Given the bijective nature of the parameterization mapping $h_{\mathcal{X}} : \Theta \to \mathcal{H}_\Theta(\mathcal{X})$ postulated in Assumption 5.4.2, it follows that there exists a unique $\boldsymbol{\theta} \in \Theta$ satisfying,

$$\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} \leq \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}') \|_{\mathcal{H}} \ \forall \boldsymbol{\theta}' \in \Theta;$$

i.e. $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}} \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ is unique. □

5.9.3 Theorem 5.6.3

Proof. In what follows, we first take some initial steps that are similar to those of Theorems 5.6.1 and 5.6.2 and then specialize the discussion to the cases of *(i)* strong unicity obtained under Assumption 5.6.3, and *(ii)* strong unicity of order α obtained under either Assumption 5.6.4 or 5.6.5. As before, Assumption 5.4.1 guarantees the formulation,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} Q_\infty^{\mathcal{D}}(\mathrm{D}_0^{\mathbf{w}}, \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}).$$

Furthermore, according to Assumption 5.5.1, $\boldsymbol{\theta}_0$ also satisfies

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| \mathrm{D}_0^{\mathbf{w}} - \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \|_{\mathcal{D}}$$

where $\| \cdot \|_{\mathcal{D}}$: $\mathcal{D}^{\mathbf{w}} \to \mathbb{R}_0^+$ is a norm defined on $\mathcal{D}^{\mathbf{w}}$ as $\| \cdot \|_{\mathcal{D}} \equiv g \circ Q_\infty^{\mathcal{D}}$ with $g : \mathbb{R} \to \mathbb{R}_0^+$ a strictly increasing function. Now given Assumption 5.4.2, we have $\| \mathrm{D}_0^{\mathbf{w}} - \mathrm{D}_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}} \|_{\mathcal{D}} \equiv \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$ by construction since $\| \cdot \|_{\mathcal{H}}$: $\mathcal{H}(\mathcal{X}) \to \mathbb{R}_0^+$ is a norm defined on $\mathcal{H}(\mathcal{X})$ according to $\| h \|_{\mathcal{H}} \equiv \| \mathrm{D}(h) \|_{\mathcal{D}} \ \forall h \in \mathcal{H}(\mathcal{X})$ and hence $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$ holds true. Finally, we split this proof into three parts and obtain the desired identifiable uniqueness of $\boldsymbol{\theta}_0$, under either Assumption

5.6.3, 5.6.4 or 5.6.5 respectively.

Part I. Let Assumption 5.6.3 hold. Then,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \parallel h_0 - h(\cdot,\boldsymbol{\theta}) \parallel_\infty .$$

Furthermore, for h_0 and $h_{\mathcal{X}}(\boldsymbol{\theta})$ satisfying the conditions of Assumption 5.6.3 we have by Lemma 5.5.7 that for every $h_0 \in \mathcal{H}(\mathcal{X})$, there exists a unique $h \in \mathcal{H}_\Theta(\mathcal{X})$ satisfying the strong unicity property, $\parallel h_0 - h' \parallel_\infty > \parallel h_0 - h \parallel_\infty + \gamma \parallel h - h' \parallel_\infty$ $\forall\, h' \in \mathcal{H}_\Theta(\mathcal{X})$, with $\gamma > 0$, thus conveniently restated as,

$$\parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty > \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \parallel_\infty + \gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{H})$$

since every element $h \in \mathcal{H}_\Theta(\mathcal{X})$ has a parametric representation of the form $h_{\mathcal{X}}(\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$. Now, clearly, $\parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty > \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \parallel_\infty + \gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty \ \Leftrightarrow \ \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty - \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \parallel_\infty > \gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty$ $\forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$, and hence, $\inf_{\boldsymbol{\theta}\in\Theta^*}[\parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty - \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \parallel_\infty] \geq \inf_{\boldsymbol{\theta}\in\Theta^*}[\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty]$ holds for any $\Theta^* \subseteq \Theta$. We now show that when $\Theta^* = \eta_0(\epsilon)^c$, then,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty - \parallel h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \parallel_\infty] \geq \inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty] > 0.$$

Indeed, note first that,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty] \ = \ \inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)}[\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty]$$

and hence that it is enough to show that,

$$\inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)}[\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty] > 0.$$

It is elementary that for every $\gamma > 0$, having,

$$\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty > c > 0 \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$$

for some $c > 0$ independent of $\boldsymbol{\theta}$, implies $\inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)} \gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty > 0$, and that, $\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty > c > 0 \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$ holds true whenever $h_{\mathcal{X}}(\eta_0(\epsilon))$ is an open set with $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$, because then, $\exists\, \delta > 0$ such that $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)$ is an open ball of radius δ centered at $h_{\mathcal{X}}(\boldsymbol{\theta}_0)$ satisfying $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta) \subseteq h_{\mathcal{X}}(\eta_0(\epsilon))$, and hence, by definition, $\parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty \geq \delta > 0 \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c$ where $S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c := \mathcal{H}_\Theta(\mathcal{X}) \backslash S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)$. This implies that, for every $\gamma > 0$,

$$\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty > c > 0 \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c$$

holds uniformly in $\boldsymbol{\theta} \in \Theta$ for every $0 < c < \delta/\gamma$. Thus, the desired result,

$$\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty] > 0$$

is implied by Assumption 5.6.6 which ensures the openess of $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$ and hence that $\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\gamma \parallel h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \parallel_\infty] > 0$. Finally, since $Q_\infty(\boldsymbol{\theta}) \equiv Q_\infty^{\mathcal{H}}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ satisfies,

$$g \circ Q_\infty^{\mathcal{H}}(h_0, h(\cdot; \boldsymbol{\theta})) \equiv \| h_0 - h(\cdot; \boldsymbol{\theta}) \|_{\mathcal{H}} \equiv \| h_0 - h(\cdot; \boldsymbol{\theta}) \|_\infty \ \forall \boldsymbol{\theta} \in \Theta$$

with strictly increasing g, it follows that,

$$\inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [\| h_0 - h(\cdot; \boldsymbol{\theta}) \|_\infty - \| h_0 - h(\cdot; \boldsymbol{\theta_0}) \|_\infty] > 0 \Leftrightarrow \inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0.$$

We thus conclude that strong unicity implies identifiable uniqueness under Assumptions 5.4.2, 5.5.1, 5.6.3 and 5.6.6, i.e. that $\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_\infty + \gamma \| h - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_\infty \ \forall \boldsymbol{\theta} \in \Theta \Rightarrow \inf_{\boldsymbol{\theta} \in \eta_0(\epsilon)^c} [Q_\infty(\boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta}_0)] > 0 \, \forall \boldsymbol{\theta} \in \Theta$.

Part II. Let Assumption 5.6.4 hold instead of 5.6.3. Then, except for some trivial minor details, the same argument holds. In particular, we now have,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$$

where $\| \cdot \|_{\mathcal{H}}$ satisfies,

$$\| h \|_{\mathcal{H}} = \Big(\int_{\mathcal{X}} |h|^p d\mu \Big)^{1/p} \ \forall h \in \mathcal{H}(\mathcal{X})$$

with $1 < p < \infty$. Since $h_0 \in L^p(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mu_{\mathcal{X}})$ with $1 < p < \infty$ and $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$ where $\mathcal{H}_\Theta(\mathcal{X})$ is a finite dimensional subset of $\mathcal{H}(\mathcal{X})$, we have by Lemma 5.5.8 that for every $h_0 \in \mathcal{H}(\mathcal{X})$, when there exists a unique best approximation $h \in \mathcal{H}_\Theta(\mathcal{X})$ to $h_0 \in \mathcal{H}(\mathcal{X})$ then it is strongly unique of order $\alpha = \max\{p, 2\}$. In other words, $\exists \, \gamma(h_0) > 0 : \| h_0 - h' \|_{\mathcal{H}} > \| h_0 - h \|_{\mathcal{H}} + \gamma \| h - h' \|_{\mathcal{H}}^\alpha \ \forall h' \in \mathcal{H}_\Theta(\mathcal{X})$. This property is conveniently restated as

$$\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}} + \gamma \| h - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^\alpha \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$$

since every element $h \in \mathcal{H}_\Theta(\mathcal{X})$ has a parametric representation of the form $h_{\mathcal{X}}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta$. The existence of an element of best approximation follows from Lemma 5.5.4 by noting that every uniformly convex normed vector space is strictly convex (Cheney (1982, p.23)). As before, the elementary step $h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}} + \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^\alpha \quad \Leftrightarrow$
$\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}} > \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^\alpha \ \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X})$
implies,

$$\inf_{\boldsymbol{\theta} \in \Theta^*} [\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}}] \geq \inf_{\boldsymbol{\theta} \in \Theta^*} [\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}}^\alpha]$$

holds for any $\Theta^* \subseteq \Theta$. Again, we are interested in the case $\Theta^* = \eta_0(\epsilon)^c$, and to obtain

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}}] > 0$$

it is enough to show that $\inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}}] > 0$. Since for every $\gamma > 0$ and $\alpha > 1$, having,

$$\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}} > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$$

for some $c > 0$ constant, implies,$\inf_{h_{\mathcal{X}}(\boldsymbol{\theta})\in h_{\mathcal{X}}(\eta_0(\epsilon)^c)} \gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}} > 0$, and that, $\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}} > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in h_{\mathcal{X}}(\eta_0(\epsilon)^c)$ holds true if $h_{\mathcal{X}}(\eta_0(\epsilon))$ is an open set satisfying $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$ for every $\epsilon > 0$, by the same argument as before. Thus, for every $\gamma > 0$ it holds true that,

$$\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}} > c > 0 \, \forall h_{\mathcal{X}}(\boldsymbol{\theta}) \in S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c$$

uniformly in $\boldsymbol{\theta} \in \Theta$, for every $0 < c < (\delta/\gamma)^{1/\alpha}$ where

$$S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta)^c := \mathcal{H}_{\Theta}(\mathcal{X}) \backslash S(h_{\mathcal{X}}(\boldsymbol{\theta}_0), \delta).$$

Hence, $\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}}] > 0$ is implied by Assumption 5.6.6 which ensures the openess of $h_{\mathcal{X}}(\boldsymbol{\theta}_0) \in h_{\mathcal{X}}(\eta_0(\epsilon))$ and hence that $\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\gamma \| h_{\mathcal{X}}(\boldsymbol{\theta}_0) - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}}] > 0$. Finally, since $Q_{\infty}(\boldsymbol{\theta}) \equiv Q^{\mathcal{H}}_{\infty}(h_0, h_{\mathcal{X}}(\boldsymbol{\theta}))$ satisfies,

$$g \circ Q^{\mathcal{H}}_{\infty}(h_0, h(\cdot;\boldsymbol{\theta})) \equiv \| h_0 - h(\cdot;\boldsymbol{\theta}) \|_{\mathcal{H}} \equiv \| h_0 - h(\cdot;\boldsymbol{\theta}) \|_{\mathcal{H}} \ \forall \boldsymbol{\theta} \in \Theta$$

with strictly increasing g, it follows that,

$$\inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[\| h_0 - h(\cdot;\boldsymbol{\theta}) \|_{\mathcal{H}} - \| h_0 - h(\cdot;\boldsymbol{\theta_0}) \|_{\mathcal{H}}] > 0 \Leftrightarrow \inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[Q_{\infty}(\boldsymbol{\theta}) - Q_{\infty}(\boldsymbol{\theta}_0)] > 0.$$

We thus conclude that strong unicity of order α implies identifiable uniqueness under Assumptions 5.4.2, 5.5.1, 5.6.4 and 5.6.6, i.e. that $\| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}) \|_{\mathcal{H}} > \| h_0 - h_{\mathcal{X}}(\boldsymbol{\theta}_0) \|_{\mathcal{H}} + \gamma \| h - h_{\mathcal{X}}(\boldsymbol{\theta}) \|^{\alpha}_{\mathcal{H}} \ \forall \boldsymbol{\theta} \in \Theta \Rightarrow \inf_{\boldsymbol{\theta}\in\eta_0(\epsilon)^c}[Q_{\infty}(\boldsymbol{\theta}) - Q_{\infty}(\boldsymbol{\theta}_0)] > 0 \, \forall \boldsymbol{\theta} \in \Theta$.

Part III. Finally, let Assumption 5.6.5 hold instead of 5.6.3 or 5.6.4. Now,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \| h_0 - h(\cdot, \boldsymbol{\theta}) \|_{\mathcal{H}}$$

where $\| \cdot \|_{\mathcal{H}}$ is such that $(\mathcal{H}(\mathcal{X}), \| \cdot \|_{\mathcal{H}})$ is a uniformly convex Banach space of power type $p > 1$. Since $h_0 \in \mathcal{H}(\mathcal{X})$ and $h_{\mathcal{X}}(\boldsymbol{\theta}) \in \mathcal{H}_{\Theta}(\mathcal{X})$ where $\mathcal{H}_{\Theta}(\mathcal{X})$ is a closed convex subspace of $\mathcal{H}(\mathcal{X})$, we have by Lemma 5.5.8 that for every $h_0 \in \mathcal{H}(\mathcal{X})$, when there exists a unique best approximation $h \in \mathcal{H}_{\Theta}(\mathcal{X})$ to $h_0 \in \mathcal{H}(\mathcal{X})$, then, it is strongly unique of order p. As we have already seen, this form of strong unicity implies the identifiable uniqueness of $\boldsymbol{\theta}_0$. The existence of an element of best approximation follows from the fact that a closed convex subset of a uniformly convex Banach space is proximinal (Cheney (1982, p.22)).

□

Chapter 6

Conclusion

This short chapter summarizes very briefly the main findings contained in this thesis and reviews its most important limitations. Finally, a number of future research directions are also proposed.

This thesis introduced a novel *sieve extremum estimator* that relies on auxiliary statistics through the principle of *indirect inference*. This estimator was designed to allow for an econometric analysis that deals well with two main problems in econometric analysis. The first is related to the restrictiveness of working with parameter spaces of finite complexity and the consequent restrictiveness of correct specifications axioms. The second is related to the possible failure of classical estimators (e.g. due to intractable criterion functions) in the presence of high-dimensional dynamic models featuring unobserved variables.

Making use of high-level assumptions, Chapter 2 introduced novel convergence rate and asymptotic distribution theorems that hold for the entire class of appropriately smooth sieve extremum estimators. As recently, pointed out by Chen (2007) such general theorems are currently unavailable. Hence, these results should add to the existing literature of sieve extremum estimation. These theorems relied on a number of novel smoothness concepts that have been introduced and characterized in Appendix C. The high-level assumptions used in Chapter 2 were useful in highlighting the conditions that ensure appropriate convergence properties for sieve extremum estimators. In particular, this allowed for a clear separation between the *general theory*, which applies to most sieve extremum estimators, and the *special theory*, that applies only to the case of *semi-nonparametric indirect inference* (SNPII) estimators.

Chapter 3 delivered primitive conditions for the measurability, consistency, convergence rate and asymptotic distribution of a special sub-class of SNPII estimators. In particular, this chapter derived the $\sqrt{T}$-consistency and asymptotic Gaussianity of SNPII estimators relying on an infinite number of parametric auxiliary statistics. Similar results were obtained for estimation of appropriate functionals of the true

parameter $\boldsymbol{\theta}_0$. This chapter offered also a characterization of statistical inference conducted using a double asymptotic approximation of the large sample distribution of the SNPII estimator.

Chapter 4 provided Monte Carlo evidence of the small-sample behavior of the SNPII estimator. This chapter suggested advantages in the use of flexible econometric techniques like SNPII estimation that deliver generality to correct specification axioms. The use of SNPII in the context of theory-driven models was also analyzed. In this respect, Chapter 4 clarified also that the SNPII framework can be used in conjunction with restrictions stemming from economic theory to guide the 'design' of econometric models. A brief note on problems of accuracy related to the normalization of random variables in simulations from dynamic models was introduced in Appendix D.

Finally, Chapter 5 revealed that the literature of *Approximation Theory* can also be used to verify if identifiable uniqueness conditions hold for a large class of extremum estimators on misspecified models. In particular, this chapter reduced the verification of identifiable uniqueness conditions to the verification of strong unicity of best approximations. By doing this, Chapter 5 offered a theory that yields identifiable uniqueness assumptions easier to verify in various contexts.

In essence, it seems fair to say that this thesis has established the basic fundamental results that allow for a future theory of semi-non parametric indirect inference estimation to be further developed. Clearly, a large number of extensions should however be pursued if we are to have a better understanding of both the advantages and limitations of this methodology. In what follows I describe some important extensions.

First, it is very easy to extend the existing results so as to accommodate for exogenous variables. In particular, an extension to dynamic models of the form considered in Gourieroux et al. (1993) is easily accomplished. The only difference is that instead of being indexed by a finite parameter vector, the unknown functions that define the dynamic equations are now allowed to be of a considerably more general nature.

Second, it should be noted that it is of great practical interest to devise tests that allow the researcher to decide whether the sieves are large enough. Indeed, there is *a priori* no reason to preclude the possibility that the 'true parameter' $\boldsymbol{\theta}_0$ lies on a early sieve $\Theta_T \subseteq \Theta$. Following Gourieroux et al. (1993), such tests can in principle be derived from the SNPII criterion function as 'correct specification tests' for any given sieve. Adoption of a sieve selection strategy (a *stopping rule* for sieve expansion) that relies on the data would however imply that the sieves are themselves random. Further theory is then needed to deal with random sieves as there are inferential problems that must be addressed. Such an extension is likely

to be relevant in applications since it allows for a data dependent sieve structure.

Third, it is important to establish optimal sieve expansion rates that minimize the estimator's variance. In this thesis we have only made use of lower bounds on the sieve expansion rates that are designed to ensure the convergence in distribution of the SNPII estimator. Results on sieve rate optimality would also allow us obtain an appropriate description of the SNPII estimator in terms of efficiency.[1]

Fourth, in what concerns convergence rates, it is also important to deliver a more complete theory of the 'indirect' restrictions that are imposed on the parameter space Θ by the assumptions on the binding function. In particular, further research should yield a more complete picture of the restrictions discussed in Section 3.9.

Fifth, it is of interest to analyze in more detail the asymptotic behavior of SNPII estimators relying on alternative auxiliary statistics. As pointed out in Chapter 2, from the outset, nothing restricts the SNPII estimator from making use of a single auxiliary estimator. The intuitive problem with this approach is however that, for the auxiliary estimator to be informative about the parameter of interest $\boldsymbol{\theta}_0$, then the auxiliary space $\mathcal{B}$ should be at least as large and complex as Θ. SNPII estimators relying on a single, multiple, or infinitely many auxiliary nonparametric or sieve estimators should thus be analyzed.

Sixt, it is certainly of interest to study further the finite sample properties of the SNPII estimator. As explained in Chapter 4, by allowing the set on which the estimator takes values to increase with sample size, we have made correct specification axioms more plausible. However, by following this trail, we have shifted our concerns from the asymptotic behavior to the finite-sample behavior of our estimator. Indeed, in finite samples, the sieve restrictions are likely to be binding, and this might result in the presence of significant finite-sample bias. Chapter 4 has provided some first limited Monte Carlo evidence of the finite-sample behavior of the SNPII estimator. As pointe out there, further research should however include *(i)* a detailed analysis of alternative criterion functions, auxiliary estimators and sieves; *(ii)* a comparison of alternative sieves ensuring the appropriate stability and fading memory properties on dynamic models and *(iii)* in the context of rational expectation models derived from economic theory, the behavior of SNPII estimators should be analyzed in conjunction with alternative *solution methods* that deliver appropriate approximation of policy functions.

Seventh, the benefits of SNPII estimation should also be analyzed in terms of its ability to deliver potentially better results in terms of *(i)* describing the nonlinear and asymmetric relation between economic variables; *(ii)* providing a more accurate

[1]The indirect inference literature suggests that the use of efficient auxiliary estimators might result in the SNPII estimator being itself efficient. However, in the context of SNPII, efficiency depends also on the sieve expansion rate.

account of the dynamic properties of data; and *(iii)* improving in-sample fit and out-of-sample forecasts.

Finally, a quite natural research topic in econometrics consists obviously of extending the present results to non-stationary unit-root data. This however, is likely to take its time as both *sieve estimation* and *indirect inference* are still essentially constrained to the stationary world.[2]

[2] An essential problem posed by non-stationary data concerns the fact that uniform convergence of the indirect inference criterion function might fail. This happens because convergence properties of auxiliary estimators for those $\boldsymbol{\theta}$ that imply non-stationarity are generally different from the convergence properties obtained for a $\boldsymbol{\theta}$ which implies stationarity. Clearly, this complicates the estimator consistency arguments (of both sieve and II estimators) based on uniform convergence of criterion functions. A possible solution consists of applying different weights to different auxiliary estimators so that, for any given $\boldsymbol{\theta}$, only a subset of 'well-behaved' auxiliary estimators 'receive' (asymptotically) positive weight.

Appendix A

Auxiliary Definitions Lemmas and Propositions

Definition A.1. (Separable Space) *A topological space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is said to be separable if and only if it has a countable dense subset.*

Definition A.2. (Polish Space) *A topological space is said to be a Polish space if it separable and there exists a metric that generates the topology for which the space is complete. Any separable complete metric space is thus a Polish space.*

Definition A.3. (Metrizable Space) *A topological space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is said to be metrizable if and only if there exists a metric* $\delta_{\mathbb{A}}$ *that induces* $\mathcal{T}_{\mathbb{A}}$ *on* $\mathbb{A}$*, i.e. such that sets of* $\mathcal{T}_{\mathbb{A}}$ *are open w.r.t.* $\delta_{\mathbb{A}}$*.*

Definition A.4. (Regular Space) *A topological space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is called regular if for every point* $a \notin \mathbb{A}_0 \subset \mathbb{A}$ *there are disjoint open sets* $\mathbb{A}_1$ *and* $\mathbb{A}_2$ *with* $a \in \mathbb{A}_1$ *and* $\mathbb{A}_0 \subset \mathbb{A}_2$*.*

Definition A.5. (Base for a Topology) *A base for a topology* $\mathcal{T}$ *is any collection* $\mathcal{T}_0 \subset \mathcal{T}$ *such that for every* $\mathcal{T}_1 \subset \mathcal{T}$*, we have* $\mathcal{T}_1 = \bigcup \left\{ \mathcal{T}_0' \in \mathcal{T}_0 : \mathcal{T}_0' \subset \mathcal{T}_1 \right\}$*.*

Definition A.6. (Second Countable Space) *A topological space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is said to be second countable if* $\mathcal{T}_{\mathbb{A}}$ *has a countable base.*

Lemma A.7. (Urysohn-Tychonoff Theorem) [Klambauer 1973, Proposition 31, p.257] *Every regular second-countable topological space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is metrizable.*

Definition A.8. (Hausdorff Space) *A topological space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is Hausdorff if and only if* $\forall (a_1, a_2) \in \mathbb{A} \times \mathbb{A}$ *there exists open sets* $\mathbb{A}_1 \subset \mathbb{A}$ *and* $\mathbb{A}_2 \subset \mathbb{A}$ *such that* $a_1 \in \mathbb{A}_1$*,* $a_2 \in \mathbb{A}_2$ *and* $A_1 \cap \mathbb{A}_2 = \emptyset$*.*

Lemma A.9. (Metrizable-Hausdorff Space) [Sutherland 2009, Proposition 11.4, p.110] *Every metrizable space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is Hausdorff.*

Definition A.10. (Measurable Map) *Let* $(\mathbb{A}, \mathfrak{B}(\mathbb{A}))$ *and* $(\mathbb{B}, \mathfrak{B}(\mathbb{B}))$ *be measurable spaces. A map* $f : \mathbb{A} \to \mathbb{B}$ *is* $\mathfrak{B}(\mathbb{B})/\mathfrak{B}(\mathbb{A})$*-measurable if* $f^{-1}(B) \in \mathfrak{B}(\mathbb{A})$ *for every* $B \in \mathfrak{B}(\mathbb{B})$.

Lemma A.11. (Measurable Map) [Billingsley (1995, Theorem 13.1, p.182)] *Let* $(\mathbb{A}, \mathfrak{A})$ *and* $(\mathbb{B}, \mathfrak{B})$ *be measurable spaces. Let* $f : \mathbb{A} \to \mathbb{B}$ *be such that* $f^{-1}(B) \in \mathfrak{A}$ *for every* $B \in \mathcal{B}_0$ *and let* $\mathfrak{B}$ *be generated by* $\mathcal{B}_0$*, then* f *is* $\mathcal{B}/\mathcal{A}$*-measurable.*

Lemma A.12. (Inverse of Continuous Operator) [Klambauer 1973, Proposition 4, p.234] *Let* $(\mathbb{A}, \mathfrak{A})$ *and* $(\mathbb{B}, \mathfrak{B})$ *be topological spaces. A map* $f : \mathbb{A} \to \mathbb{B}$ *is continuous map if and only if its inverse is an open map.*

The following is an immediate Corollary of Lemmas A.11 and A.12.

Corollary A.13. (Continuous Borel Map) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological spaces with Borel* σ*-algebra* $\mathfrak{B}(\mathbb{A})$ *and* $\mathfrak{B}(\mathbb{B})$ *generated by* $\mathcal{T}_{\mathbb{A}}$ *and* $\mathcal{T}_{\mathbb{B}}$ *respectively. Then a continuous map* $f : \mathbb{A} \to \mathbb{B}$ *is* $\mathfrak{B}(\mathbb{B})/\mathfrak{B}(\mathbb{A})$*-measurable.*

Lemma A.14. (Measurable Composition) [Billingsley 1995, Theorem 13.1, p.182] *Let* $(\mathbb{A}, \mathfrak{A})$*,* $(\mathbb{B}, \mathfrak{B})$ *and* $(\mathbb{C}, \mathfrak{C})$ *be measurable spaces. Let* $f : \mathbb{A} \to \mathbb{B}$ *be* $\mathcal{B}/\mathcal{A}$*-measurable and* $g : \mathbb{B} \to \mathbb{C}$ *be* $\mathfrak{C}/\mathfrak{B}$*-measurable. Then* $g \circ f : \mathbb{A} \to \mathbb{C}$ *is* $\mathfrak{C}/\mathfrak{A}$*-measurable.*

Lemma A.15. (Continuous Projections) [Gamelin and Greene 1999, Theorem 12.1, p.101] *Let* $(\mathbb{A}_i, \mathcal{T}_{\mathbb{A}_i})$ *be topological spaces for all* i *in some set* $\mathbb{I}$ *and let* $\mathbb{A} = \times_{i \in \mathbb{I}} \mathbb{A}_i$*. The product topology is the smallest topology making the coordinate projections* $\pi_i : \mathbb{A} \to \mathbb{A}_i$ *continuous* $\forall\, i \in \mathbb{I}$.

The following result thus follows as a Corollary of Lemma A.15.

Corollary A.16. (Product Topology Convergent Sequences) [James 1987, Corollary 2.12, p.33] *Let* $(\mathbb{A}_i, \mathcal{T}_{\mathbb{A}_i})$ *be topological spaces for all* i *in some set* $\mathbb{I}$ *and let* $\mathbb{A} = \times_{i \in \mathbb{I}} \mathbb{A}_i$*. Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *be the product space with product topology* $\mathcal{T}_{\mathbb{A}}$*. A sequence* $\{a_n\}_{n \in \mathbb{N}} \subset \mathbb{A}$ *satisfies* $a_n \to a \in \mathbb{A}$ *if and only if* $\pi_i(a_n) \to \pi_i(a) \in \mathbb{A}_i\ \forall\, i \in \mathbb{I}$.

Also, a Corollary of Lemmas A.13 and A.15 is as follows.

Corollary A.17. (Measurable Projections) *Let* $(\mathbb{A}_i, \mathcal{T}_{\mathbb{A}_i})$ *be topological spaces with Borel* σ*-algebra* $\mathfrak{B}(\mathbb{A}_i)$ *generated by* $\mathcal{T}_{\mathbb{A}_i}$ *for all* i *in some set* $\mathbb{I}$*. Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *be the product space* $\mathbb{A} = \times_{i \in \mathbb{I}} \mathbb{A}_i$ *with product topology* $\mathcal{T}_{\mathbb{A}}$ *and Borel* σ*-algebra* $\mathfrak{B}(\mathbb{A})$ *generated by* $\mathcal{T}_{\mathbb{A}}$*. Then the projection maps* $\pi_i : \mathbb{A} \to \mathbb{A}_i$ *are* $\mathfrak{B}(\mathbb{A}_i)/\mathfrak{B}(\mathbb{A})$*-measurable* $\forall\, i \in \mathbb{I}$.

Lemma A.18. (Continuous Map into Product Spaces) [Gamelin and Greene 1999, Theorem 12.2, p.101] *Let $(\mathbb{A}_i, \mathcal{T}_{\mathbb{A}_i})$ be topological spaces for all i in some set $\mathbb{I}$ and let $\mathbb{A} = \times_{i\in\mathbb{I}}\mathbb{A}_i$. Let $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ be the product space with product topology $\mathcal{T}_{\mathbb{A}}$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ be some topological space. An operator $f : \mathbb{B} \to \mathbb{A}$ is continuous at $b \in \mathbb{B}$ if and only if $\pi_i \circ f : \mathbb{B} \to \mathbb{A}_i$ is continuous at b for every $i \in \mathbb{I}$.*

Lemma A.19. (Tychonoff's Theorem) [Dudley 2002, Theorem 2.2.8, p.39] *Let $(\mathbb{A}_i, \mathcal{T}_i)$ be compact topological spaces for each i in a set $\mathbb{I}$. Then the Cartesian product $\times_{i\in\mathbb{I}}\mathbb{A}_i$ with product topology is compact.*

Lemma A.20. (Subsets and Countable Products of Regular Spaces) [Munkres 2000, Theorem 31.2, p.196] *Any subspace of a regular space is regular. Any product of regular spaces is regular.*

Lemma A.21. (Countable Products of Separable Spaces) [Davidson 1994, Theorem 6.16, p.103] *Let $(\mathbb{A}_i, \mathcal{T}_{\mathbb{A}_i})$ be topological spaces for all i in some countable set $\mathbb{I}$ and let $\mathbb{A} = \times_{i\in\mathbb{I}}\mathbb{A}_i$. Let $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ be the product space with product topology $\mathcal{T}_{\mathbb{A}}$. Then $\mathbb{A}$ is separable if and only if $\mathbb{A}_i$ is separable for every $i \in \mathbb{I}$.*

Lemma A.22. (Metrization of Product Topology) [Dudley 2002, Proposition 2.4.4, p.50] *For every sequence of metric spaces $\left\{(\mathbb{A}_i, \delta_{\mathbb{A}_i})\right\}_{i\in\mathbb{N}}$, the topological product space $(\times_{i\in\mathbb{N}}\mathbb{A}_i, \mathcal{T}_{\mathbb{A}})$ with product topology $\mathcal{T}_{\mathbb{A}}$ is metrizable by the product-metric,*

$$\delta_{\mathbb{A}}(a, a') := \sum_{i\in\mathbb{N}} \frac{1}{2^i} \frac{\delta_{\mathbb{A}_i}(a_i, a_i')}{1 + \delta_{\mathbb{A}_i}(a_i, a_i')} \quad \forall\ (a, a') = \left(\{a_i\}_{i\in\mathbb{N}}, \{a_i'\}_{i\in\mathbb{N}}\right) \in \mathbb{A} \times \mathbb{A}.$$

Remark A.23. *Uncountable product spaces with product topology are not metrizable.*

Lemma A.24. (Algebra on Product Spaces) [Dudley 2002, Proposition 4.1.7, p.119] *Let $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ be any two topological spaces. Let $(\mathbb{A} \times \mathbb{B}, \mathcal{T}_{\mathbb{A}\times\mathbb{B}})$ be the product space with product Tychonoff's topology $\mathcal{T}_{\mathbb{A}\times\mathbb{B}}$ and let $\mathfrak{B}(\mathbb{A} \times \mathbb{B})$ denote the Borel σ-algebra generated by the product topology $\mathcal{T}_{\mathbb{A}\times\mathbb{B}}$ on $\mathbb{A}\times\mathbb{B}$. Then $\mathfrak{B}(\mathbb{A}\times\mathbb{B})$ includes the product σ-algebra $\mathfrak{B}(\mathbb{A}) \otimes \mathfrak{B}(\mathbb{B})$. If both $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ are second-countable then the two σ-algebras on $\mathbb{A} \times \mathbb{B}$ are equal.*

Lemma A.25. (Separability and Second-Countability) [Dudley 2002, Proposition 2.1.4, P.31] *A metric space $(\mathbb{A}, \delta_{\mathbb{A}})$ is separable if and only if it is second-countable.*

Lemma A.26. (Measurable Maps and Product σ-Algebra) [Foland 2009, p.24] *Let $(\mathbb{A}, \mathfrak{A})$ and $(\mathbb{B}_i, \mathfrak{B}_i)$ be measurable spaces for all i in some set $\mathbb{I}$. Let $(\mathbb{B}, \mathfrak{B})$ be the product space $\mathbb{B} = \times_{i\in\mathbb{I}}\mathbb{B}_i$ with product σ-algebra $\mathfrak{B} = \otimes_{i\in\mathbb{I}}$. Then the map $f : \mathbb{A} \to \mathbb{B}$ is $\mathfrak{B}/\mathfrak{A}$-measurable if and only if the projection maps $\pi_i \circ f : \mathbb{A} \to \mathbb{B}_i$ are $\mathfrak{B}_i/\mathfrak{A}$-measurable $\forall i \in \mathbb{I}$.*

The following is obtained as a Corollary of Lemmas A.24 and A.26.

Corollary A.27. (Measurable Maps into Product Spaces) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ *be topological spaces with Borel* σ*-algebra* $\mathfrak{B}(\mathbb{A})$ *and* $\mathfrak{B}(\mathbb{B}_i)$ *generated by* $\mathcal{T}_{\mathbb{A}}$ *and* $\mathcal{T}_{\mathbb{B}_i}$ *respectively for all* i *in some set* $\mathbb{I}$*. Let* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be the product space* $\mathbb{B} = \times_{i \in \mathbb{I}} \mathbb{B}_i$ *with product topology* $\mathcal{T}_{\mathbb{B}}$ *and Borel* σ*-algebra* $\mathfrak{B}(\mathbb{B})$ *generated by* $\mathcal{T}_{\mathbb{B}}$*. Then the map* $f : \mathbb{A} \to \mathbb{B}$ *is* $\mathfrak{B}(\mathcal{T}_{\mathbb{B}})/\mathfrak{B}(\mathcal{T}_{\mathbb{A}})$*-measurable if the projection maps* $\pi_i \circ f : \mathbb{A} \to \mathbb{B}_i$ *are* $\mathfrak{B}(\mathbb{B}_i)/\mathfrak{B}(\mathbb{A})$*-measurable* $\forall i \in \mathbb{I}$*.*

Definition A.28. (Topological Vector Space) *A topological vector space* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is a vector space* $\mathbb{A}$ *endowed with a topology* $\mathcal{T}_{\mathbb{A}}$ *such that vector addition and scalar multiplication are continuous functions.*

Lemma A.29. (Continuous Composition) [Sutherland 2009, Proposition 8.4, p.84] *Let* $(\mathbb{A}, \delta_{\mathbb{A}})$*,* $(\mathbb{B}, \delta_{\mathbb{B}})$ *and* $(\mathbb{C}, \delta_{\mathbb{C}})$ *be topological spaces and* $f : \mathbb{A} \to \mathbb{B}$ *and* $g : \mathbb{B} \to \mathbb{C}$ *be continuous at* $a \in \mathbb{A}$ *and* $b \in \mathbb{B}$ *respectively. Then* $g \circ f : \mathbb{A} \to \mathbb{C}$ *is continuous at* $a \in \mathbb{A}$*.*

Lemma A.30. (Measurable Maps) [Klein and Thompson 1984, Lemma 13.2.3,p.154] *Let* $(\Omega, \mathcal{F})$ *be a measurable space and* $(\Theta, \delta_{\Theta})$ *be a separable metric space. If* $Q(\omega, \cdot) : \Theta \to \mathbb{R}_0^+$ *is continuous in* Θ *for every* $\omega \in \Omega$ *and* $Q(\cdot, \boldsymbol{\theta}) : \Omega \to \mathbb{R}_0^+$ *is measurable for every* $\boldsymbol{\theta} \in \Theta$*, then* $Q : \Omega \times \Theta \to \mathbb{R}_0^+$ *is* $\mathcal{F} \times \mathfrak{B}(\Theta)$*-measurable.*

Lemma A.31. (Measurability) [Debreu 1967, Theorem 4.5] *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $(\mathbb{A}, \delta_{\mathbb{A}})$ *be a complete separable metric space with Borel* σ*-algebra* $\mathfrak{B}(\mathbb{A})$*. Let the random sieve-correspondence* $\mathbb{A}_T : \Omega \to \mathbb{A}$ *have a measurable graph* $\mathfrak{gr}(\mathbb{A}_T) \in \mathcal{F} \otimes \mathfrak{B}(\mathbb{A})$ *and the sieves* $\mathbb{A}_T(\omega) \subset \mathbb{A}$ *are non-empty and compact for every* $\omega \in \Omega$*. Finally, let the criterion mapping* $f_T : \mathfrak{gr}(\mathbb{A}_T) \to \mathbb{R}_0^+$ *be* $\mathcal{F} \otimes \mathfrak{B}(\mathbb{A})$*-measurable and* $f_T(\omega) : \mathbb{A} \to \mathbb{R}_0^+$ *be continuous on* $\mathbb{A}$*. Then* $f_T^{\inf} : \Omega \to \mathbb{R}_0^+$ *is* $\mathcal{F}^P/\mathfrak{B}(\mathbb{R}_0^+)$*-measurable and the minimizer set* $\widehat{\mathbb{A}\Omega} \in \mathbb{A} \times \Omega$ *defined as* $\hat{\mathbb{A}} : \Omega \to \Theta$ *satisfying* $\hat{\mathbb{A}}(\omega) := \left\{ a \in \mathbb{A}_T(\omega) : f_T(\omega, a) = \inf_{\boldsymbol{\theta} \in \mathbb{A}_T(\omega)} f_T(\omega, a) \right\}$ *for every* $\omega \in \Omega$ *belongs to* $\mathcal{F}^P \otimes \mathfrak{B}(\Theta)$*.*

Lemma A.32. (Measurable Selection) [Hildenbrand 1974, p.55] *Let* $(\Omega, \mathcal{F})$ *be a measurable space and* $\mathbb{A}$ *a complete separable metric space with its Borel* σ*-field* $\mathfrak{B}(\mathbb{A})$ *and* $f_T^{\sup} : \Omega \to 2^{\mathbb{A}}$ *a closed valued correspondence s.t.* $\{\omega \in \Omega : f_T^{\sup}(\omega) \cap \mathbb{A}^*\} \in \mathcal{F}$ *for every closed subset* $\mathbb{A}^* \subset \mathbb{A}$*. Then* $f_T^{\sup} : \Omega \to 2^{\mathbb{A}}$ *admits a measurable selector, i.e. there exists a map* $\hat{a}_T : \Omega \to 2^{\mathbb{A}}$ *that is measurable and for every* $\omega \in \Omega$ *it satisfies* $\hat{a}_T(\omega) \in f_T^{\sup}(\omega)$*.*

Corollary A.33. (Measurable Extrema) [White and Wooldrige 1991, Theorem 2.2, p.646] *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a complete probability space and* $(\mathbb{A}, \delta_{\mathbb{A}})$ *be a complete separable metric space. Let* $\{\mathbb{A}_n\}_{n \in \mathbb{N}}$ *be a sequence of compact subsets of* $\mathbb{A}$*. Let* f_n :

$\Omega \times \mathbb{A}_n \to \mathbb{R}$ *be* $\mathcal{F} \otimes \mathfrak{B}(\mathbb{A}_n)/\mathfrak{B}(\mathbb{R})$*-measurable for every* $n \in \mathbb{N}$ *and* $f_n(\omega, \cdot) : \mathbb{A} \to \mathbb{R}$ *be continuous on* $\mathbb{A}_n$ *for every* $(\omega, n) \in \Omega \times \mathbb{N}$*. Then there exists an* $\mathcal{F}/\mathfrak{B}(\mathbb{A}_n)$*-measurable map* $\hat{a}_n : \Omega \to \mathbb{A}_n$ *satisfying* $f_n(\omega, \hat{a}_n(\omega)) = \inf_{a \in \mathbb{A}_n} f_n(\omega, a)$ *for every* $\omega \in \Omega$ *and every* $n \in \mathbb{N}$*.*

Definition A.34. (Metric Equivalence) *Let* $\mathbb{A}$ *be a set. Two metrics,* $\delta^1_{\mathbb{A}} : \mathbb{A} \times \mathbb{A} \to \mathbb{R}$ *and* $\delta^2_{\mathbb{A}} : \mathbb{A} \times \mathbb{A} \to \mathbb{R}$ *are said to be topologically equivalent if they define the same open sets, i.e. if they induce the same topology* $\mathcal{T}_{\mathbb{A}}$ *on* $\mathbb{A}$*.*

Remark A.35. *Let* $\delta^1_{\mathbb{A}}$ *and* $\delta^2_{\mathbb{A}}$ *be any two topologically equivalent metrics on the set* $\mathbb{A}$*. If a sequence in* $\mathbb{A}$ *is* $\delta^1_{\mathbb{A}}$*-convergent then it is also* $\delta^2_{\mathbb{A}}$*-convergent.*

Definition A.36. (Lipschitz Stronger/Weaker Metric) *Given a pair of metrics* $\delta_{\mathbb{A}}$ *and* $\delta'_{\mathbb{A}}$ *defined on the product* $\mathbb{A} \times \mathbb{A}$ *of some set* $\mathbb{A}$*, the metric* $\delta_{\mathbb{A}}$ *is said to be Lispchitz weaker than* $\delta'_{\mathbb{A}}$ *if* $\exists k \in \mathbb{R}^+$ *such that* $\delta_{\mathbb{A}}(a, a') \leq k \cdot \delta'_{\mathbb{A}}(a, a') \, \forall (a, a') \in \mathbb{A} \times \mathbb{A}$*. The metric* $\delta'_{\mathbb{A}}$ *is also said to be Lispchitz stronger than* $\delta_{\mathbb{A}}$*. Furthermore, if* $\exists (k, k') \in \mathbb{R}^+ \times \mathbb{R}^+$ *such that* $k \cdot \delta'_{\mathbb{A}}(a, a') \leq \delta_{\mathbb{A}}(a, a') \leq k' \cdot \delta'_{\mathbb{A}}(a, a') \, \forall (a, a') \in \mathbb{A} \times \mathbb{A}$ *then* $\delta_{\mathbb{A}}$ *and* $\delta'_{\mathbb{A}}$ *are said to be Lipschitz equivalent.*

Lemma A.37. (Lispchitz Topological Equivalence) [Sutherland 2009, Proposition 6.34, p.70] *A pair of Lipschitz equivalent metrics* $\delta_{\mathbb{A}}$ *and* $\delta'_{\mathbb{A}}$ *defined on the product* $\mathbb{A} \times \mathbb{A}$ *of some set* $\mathbb{A}$ *is also topologically equivalent.*

Definition A.38. (Uniform Product Metric) *Given metric spaces* $(\mathbb{A}_i, \delta_{\mathbb{A}_i})$*,* $i \in \mathbb{I}$ *where* $\mathbb{I}$ *is a countable index set and a product space* $\mathbb{A} := \times_{i \in \mathbb{I}} \mathbb{A}_i$*. The product metric* $\delta_{\mathbb{A}}(a, a') := \sup_{i \in \mathbb{I}} \delta_{\mathbb{A}_i}(a_i, a'_i) \, \forall (a, a') \in \mathbb{A} \times \mathbb{A}$ *is called the uniform product metric on* $\mathbb{A}$*.*

Proposition A.39. (Lipschitz Weaker Metrics) *Both product metrics in (3.1) are Lipschitz weaker than the uniform product metric.*

Proof. Immediate from the definitions in (3.1) since,

$$\begin{aligned}\delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}') &= \sum_{i=1}^{\infty} \frac{1}{2^i} \frac{\delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i)}{1 + \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i)} \leq \sum_{i=1}^{\infty} \frac{1}{2^i} \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i) \\ &\leq \sum_{i=1}^{\infty} \frac{1}{2^i} \sup_{i \in \mathbb{N}} \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i) = 2 \sup_{i \in \mathbb{N}} \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i).\end{aligned}$$

and also,

$$\delta_{\mathcal{B}}(\boldsymbol{\beta}, \boldsymbol{\beta}') = \sup_{i \in \mathbb{N}} \frac{1}{i} \frac{\delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i)}{1 + \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i)} \leq \sup_{i \in \mathbb{N}} \frac{1}{i} \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i) \leq \sup_{i \in \mathbb{N}} \delta_{\mathcal{B}_i}(\boldsymbol{\beta}_i, \boldsymbol{\beta}'_i), \tag{A.1}$$

□

Definition A.40. (Difference Metric) *Given a vector space $\mathbb{A}$, a metric $\delta_{\mathbb{A}}$ on $\mathbb{A}\times\mathbb{A}$ is called a difference metric if, for every pair $(a,a')\in\mathbb{A}\times\mathbb{A}$, it satisfies $\delta_{\mathbb{A}}(a,a') = \delta_{\mathbb{A}}(a-a',0_{\mathbb{A}})$ where $0_{\mathbb{A}}$ denotes the zero element of $\mathbb{A}$. For convenience, $\delta_{\mathbb{A}}(a-a',0_{\mathbb{A}})$ shall be often denoted $\delta_{\mathbb{A}}(a-a')$.*

Remark A.41. *It is easy to verify that both product metrics in (3.1) are difference metrics.*

Definition A.42. (Asymptotically Homogeneous Metric) *Given a vector space $\mathbb{A}$ and a scalar $t\in\mathbb{R}$, a difference metric $\delta_{\mathbb{A}}$ on $\mathbb{A}\times\mathbb{A}$ is said to be asymptotically homogeneous if $\delta_{\mathbb{A}}(ta) = O(t)$ as $t\to 0$, i.e. if $\lim_{t\to 0}\delta_{\mathbb{A}}(ta)/t = O(1)$ uniformly over $a\in\mathbb{A}$.*

Remark A.43. *Any norm $\|\cdot\|_{\mathbb{A}}$ on a vector space $\mathbb{A}$ is homogeneous of first degree, i.e. $\|ta\|_{\mathbb{A}} = t\|a\|_{\mathbb{A}}$ for every scalar $t\in\mathbb{R}$ and $a\in\mathbb{A}$. In general, a metric $\delta_{\mathbb{A}}$ on a vector space $\mathbb{A}\times\mathbb{A}$ does not have to satisfy a homogeneity property such as $\delta_{\mathbb{A}}(ta,ta') = t\delta_{\mathbb{A}}(a,a')$. Indeed, the product metrics introduced in (3.1) do not satisfy this property. Nonetheless, it is easy to verify that they do satisfy the asymptotic homogeneity condition introduced in Definition A.42.*

Lemma A.44. (Inverse Bijection) [Sutherland 2009, Proposition 3.18, p.13] *Let $\mathbb{A}$ and $\mathbb{B}$ be two sets and $f:\mathbb{A}\to\mathbb{B}$. The map f is invertible if and only if it is bijective.*

Definition A.45. (Homeomorphism) *Let $(\mathbb{A},\mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B},\mathcal{T}_{\mathbb{B}})$ be topological spaces. A map $f:\mathbb{A}\to\mathbb{B}$ is said to be a Homeomorphism iff it is continuous, bijective, and has continuous inverse f^{-1}.*

Lemma A.46. (Product Homeomorphisms) [Lee 2000, Proposition 3.13, p.51, James 1987, p.31] *Let $(\mathbb{A},\mathcal{T}_{\mathbb{A}})$ and $\{(\mathbb{B}_i,\mathcal{T}_{\mathbb{B}})\}_{i\in\mathbb{I}}$ be topological spaces and $\mathbb{I}$ be an arbitrary set. Let $(\mathbb{B},\mathcal{T}_{\mathbb{B}})$ denote the product space $\mathbb{B} := \times_{i\in\mathbb{I}}\mathbb{B}_i$ with Tychonoff's topology $\mathcal{T}_{\mathbb{B}}$. A map $f:\mathbb{A}\to\mathbb{B}$ is a homeomorphism if every projection map $\pi_i f:\mathbb{A}\to\mathbb{B}_i$ is a homeomorphism for every $i\in\mathbb{I}$.*

Lemma A.47. (Open Sets in Product Topology) [Basener 1973, p.13] *Let $(\mathbb{A}_i,\mathcal{T}_{\mathbb{A}_i})$ be topological spaces for every i in some set $\mathbb{I}$ and $(\mathbb{A},\mathcal{T}_{\mathbb{A}})$ be the product space $\mathbb{A} = \times_{i\in\mathbb{I}}\mathbb{A}_i$ with product topology $\mathcal{T}_{\mathbb{A}}$. Then a subset $\mathcal{O}\subseteq\mathbb{A}$ is open if and only if $\pi_i(\mathbb{O})\subseteq\mathbb{A}_i$ is open for every $i\in\mathbb{I}$.*

Proposition A.48. (Homeomorphisms with Product Topology) *Let $(\mathbb{A},\mathcal{T}_{\mathbb{A}})$ and $\{(\mathbb{B}_i,\mathcal{T}_{\mathbb{B}})\}_{i\in\mathbb{I}}$ be topological spaces and $\mathbb{I}$ be a countable set. Let $(\mathbb{B},\mathcal{T}_{\mathbb{B}})$ denote the product space $\mathbb{B} := \times_{i\in\mathbb{I}}\mathbb{B}_i$ with Tychonoff's topology $\mathcal{T}_{\mathbb{B}}$. Let $\{f_i\}_{i\in\mathbb{I}}$ denote a collection of maps $f_i:\mathbb{A}\to\mathbb{B}_i$ such that* (i) *f_i is continuous on $\mathbb{A}\,\forall i\in\mathbb{I}$;* (ii) *$f_i$ is open*

$\forall\, i \in \mathbb{I}$*; and* (iii) *for every pair* $(a, a') \in \mathbb{A} \times \mathbb{A}$, $\exists\, i \in \mathbb{I} : f_i(a) \neq f_i(a')$. *Then the product map* $f : \mathbb{A} \to \mathbb{B}$ *satisfying* $f(a) = (f_{i_1}(a), f_{i_2}(a), ...)$ *is a homeomorphism on its range.*

Proof. Continuity of the product map f follows by continuity of each projection map f_i by Lemma A.18. Openess of f follows by noting that the image $f(\mathbb{A}_{\mathcal{T}})$ of an open set $\mathbb{A}_{\mathcal{T}} \in \mathcal{T}_{\mathbb{A}}$ must be an open subset $f(\mathbb{A}_{\mathcal{T}}) \subseteq \mathbb{B}$ by Lemma A.47 since $f_i(\mathbb{A}_{\mathcal{T}}) \subseteq \mathbb{B}_i$ is an open set (i.e. $f_i(\mathbb{A}_{\mathcal{T}}) \in \mathcal{T}_{\mathbb{B}_i}$) for every $i \in \mathbb{I}$. The injective nature follows easily since, by contradiction, if $\exists\, (a, a') \in \mathbb{A} \times \mathbb{A}$ such that $f(a) = f(a')$, then by construction is must be that $f_i(a) = f_i(a')\, \forall\, i \in \mathbb{I}$, but this contradicts the assumption that for every pair $(a, a') \in \mathbb{A} \times \mathbb{A}$, $\exists\, i \in \mathbb{I} : f_i(a) \neq f_i(a')$. □

Corollary A.49. (Injective Maps in Product Spaces) *Let* $\mathbb{A}$ *and* $\{\mathbb{B}_i\}_{i \in \mathbb{I}}$ *be sets and* $\mathbb{I}$ *be a countable set. Let* $\mathbb{B}$ *denote the Cartesian product* $\mathbb{B} := \times_{i \in \mathbb{I}} \mathbb{B}_i$. *Let* $\{f_i\}_{i \in \mathbb{I}}$ *denote a collection of injective maps* $f_i : \mathbb{A} \to \mathbb{B}_i$. *Then the product map* $f : \mathbb{A} \to \mathbb{B}$ *is injective.*

Proposition A.50. (Linear Coordinate Projections) *Let* $\mathbb{B}_i$ *be a vector space for every* i *in some countable set* $\mathbb{I}$ *and* $\mathbb{B} := \times_{i \in \mathbb{I}} \mathbb{B}_i$ *be the associated product space. Then the coordinate projections* $\pi_i : \mathbb{B} \to \mathbb{B}_i$ *are linear for every* $i \in \mathbb{I}$.

Proof. Immediate since given a scalar $c \in \mathbb{R}$ and a vectors $b = (b_1, b_2, ...) \in \mathbb{B}$, the ith projection π_i satisfies $\pi_i(c \cdot b) = \pi_i((c \cdot b_1, c \cdot b_2, ...)) = c \cdot b_i$ and $c \cdot \pi_i(b) = c \cdot b_i$ and thus $\pi_i(c \cdot b) = c \cdot \pi_i(b)\, \forall\, (c, b, i) \in \mathbb{R} \times \mathbb{B} \times \mathbb{I}$. Furthermore, given a pair of vectors $(b, b') \in \mathbb{B} \times \mathbb{B}$, $\pi_i(b + b') = \pi_i((b_1 + b'_1, b_2 + b'_2, ...)) = b_i + b'_i$ and $\pi_i(b) + \pi_i(b') = b_i + b'_i$ and thus $\pi_i(b + b') = \pi_i(b) + \pi_i(b')\, \forall\, (b, b', i) \in \mathbb{B} \times \mathbb{B} \times \mathbb{I}$. □

Definition A.51. (Divergence) *Let* $\mathbb{A}$ *be a non-empty set and* $f : \mathbb{A} \times \mathbb{A} \to \mathbb{R}$. *The real-valued map* f *is said to be a divergence on* $\mathbb{A}$ *if and only if it satisfies* (i) *non-negativity* $f(a, a') \geq 0\, \forall\, (a, a') \in \mathbb{A} \times \mathbb{A}$, *and* (ii) *identity of indiscernibles* $f(a, a') = 0$ *iff* $a = a'$, $\forall\, (a, a') \in \mathbb{A} \times \mathbb{A}$.

Definition A.52. (Identifiably Unique Minimizer) *Let* $(\mathbb{A}, \delta_{\mathbb{A}})$ *be a metric space and* $f : \mathbb{A} \to \mathbb{R}$ *be some real-valued map. Then* $a_0 \in \mathbb{A}$ *is called an identifiably unique minimizer of* f *if and only if* $\inf_{a \in S^c_{a_0}(\epsilon)} |f(a) - f(a_0)| > 0$ *for every* $\epsilon > 0$.

Definition A.53. (g-Homogeneous Function) *Let* $\mathbb{A}$ *be a vector space. A function* $f : \mathbb{A} \to \mathbb{R}$ *is called* g*-homogeneous if and only if there exists a function* $g : \mathbb{A} \to \mathbb{R}$ *satisfying* $g(a_n) = O_p(1)$ *for every sequence* $\{a_n\}_{n \in \mathbb{N}}$ *satisfying* $a_n = O_p(1)$ *and* $\limsup_{n \in \mathbb{N}} g(a_n) < \infty$ *a.s. for every sequence* $\{a_n\}_{n \in \mathbb{N}}$ *satisfying* $\limsup_{n \in \mathbb{N}} a_n < \infty$ *a.s. , such that* $f(a \cdot a') = g(a) \cdot f(a')$.

Lemma A.54. (Extended Continuous Mapping Theorem) [van der Vaart and Wellner 1996, Theorem 1.11.1, p.67] *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a complete probability space and* $(\mathbb{A}, \mathfrak{B}(\mathbb{A}))$ *and* $(\mathbb{B}, \mathfrak{B}(\mathbb{B}))$ *be measurable spaces with Borel* σ*-algebras* $\mathfrak{B}(\mathbb{A})$ *and* $\mathfrak{B}(\mathbb{B})$ *respectively. Let* $f_T : \mathbb{A}_T \to \mathbb{B}$ *be measurable maps defined on subsets* $\mathbb{A}_T \subset \mathbb{A}\ \forall T \in \mathbb{N}$ *satisfying* $f_T(a_T) \to f(a)$ *for every* $a_T \to a$ *with* $a_T \in \mathbb{A}_T\ \forall T \in \mathbb{N}$, $a \in \mathbb{A}_0$ *and some measurable* $f : \mathbb{A}_0 \to \mathbb{B}$ *with* $\mathbb{A}_0 \subset \mathbb{A}$. *Let* $X_T : \Omega \to \mathbb{A}_T$ *be* $\mathcal{F}/\mathfrak{B}(\mathbb{A}_T)$*-measurable maps taking values in* $\mathbb{A}_T$ *and* X *be* $\mathcal{F}/\mathfrak{B}(\mathbb{A})$*-measurable and separable and take values in* $\mathbb{A}_0$. *Then, (i)* $X_T \xrightarrow{d} X$ *implies* $f_T(X_T) \xrightarrow{d} f(X)$, *(ii)* $X_T \xrightarrow{p} X$ *implies* $f_T(X_T) \xrightarrow{p} f(X)$, *and (iii)* $X_T \xrightarrow{a.s.} X$ *implies* $f_T(X_T) \xrightarrow{a.s.} f(X)$.

Corollary A.55. (Continuous Mapping Theorem) *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a complete probability space and* $(\mathbb{A}, \mathfrak{B}(\mathbb{A}))$ *and* $(\mathbb{B}, \mathfrak{B}(\mathbb{B}))$ *be measurable spaces with Borel* σ*-algebras* $\mathfrak{B}(\mathbb{A})$ *and* $\mathfrak{B}(\mathbb{B})$ *respectively. Let* $g : \mathbb{A} \to \mathbb{B}$ *be continuous. Let* $X : \Omega \to \mathbb{A}$ *be* $\mathcal{F}/\mathfrak{B}(\mathbb{A})$*-measurable and separable and take values in* $\mathbb{A}$. *Then, (i)* $X_T \xrightarrow{d} X$ *implies* $f(X_T) \xrightarrow{d} f(X)$, *(ii)* $X_T \xrightarrow{p} X$ *implies* $f(X_T) \xrightarrow{p} f(X)$, *and (iii)* $X_T \xrightarrow{a.s.} X$ *implies* $f(X_T) \xrightarrow{a.s.} f(X)$.

Lemma A.56. (Squeeze Theorem) [Davidson and Donsig 2009, Theorem 2.4.6, p.17] *Let* $\{a_n\}_{n\in\mathbb{N}}$, $\{b_n\}_{n\in\mathbb{N}}$ *and* $\{c_n\}_{n\in\mathbb{N}}$ *be sequences satisfying* $a_n \leq b_n \leq c_n\ \forall n \in \mathbb{N}$ *and* $\lim_{n\to\infty} a_n = b$ *and* $\lim_{n\to\infty} c_n = b$. *Then* $\lim_{n\to\infty} b_n = b$.

Proposition A.57. (Uniform Continuity Preserves Uniform Convergence) *Consider the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and let* $(\mathbb{A}, \delta_{\mathbb{A}})$, $(\mathbb{B}, \delta_{\mathbb{B}})$ *and* $(\mathbb{C}, \delta_{\mathbb{C}})$ *be measurable metric spaces with Borel* σ*-algebras* $\mathfrak{B}(\mathbb{A})$, $\mathfrak{B}(\mathbb{B})$ *and* $\mathfrak{B}(\mathbb{C})$ *respectively. Let* $g : \mathbb{B} \to \mathbb{C}$ *be a uniformly continuous map on* $\mathbb{B}$. *If* $\{f_T\}_{T\in\mathbb{N}}$ *are measurable maps* $f_T : \Omega \times \mathbb{A} \to \mathbb{B}$ *satisfying* $\sup_{a\in\mathbb{A}} \delta_{\mathbb{B}}(f_T(a), f_\infty(a)) \xrightarrow{p} 0$ *for some measurable* $f_\infty : \mathbb{A} \to \mathbb{B}$, *then* $\sup_{a\in\mathbb{A}} \delta_{\mathbb{C}}(g \circ f_T(a), g \circ f_\infty(a)) \xrightarrow{p} 0$, *and if* $\sup_{a\in\mathbb{A}} \delta_{\mathbb{B}}(f_T(a), f_\infty(a)) \xrightarrow{a.s.} 0$ *then* $\sup_{a\in\mathbb{A}} \delta_{\mathbb{C}}(g \circ f_T(a), g \circ f_\infty(a)) \xrightarrow{a.s.} 0$.

Proof. By uniform continuity of g on Y we have that for every $(\omega, T) \in \Omega \times \mathbb{N}$ and every $\epsilon > 0$, $\exists\ \epsilon' > 0$ such that having

$$\delta_Y(f_T(\omega, a), f_\infty(a)) < \epsilon' \quad \text{implies} \quad \delta_Z(g \circ f_T(\omega, a), g \circ f_\infty(\omega, a)) < \epsilon. \tag{A.2}$$

Now, convergence in probability follows since for every $T \in \mathbb{N}$ it holds true that

$$\mathbb{P}(\sup_{a\in\mathbb{A}} \delta_{\mathbb{C}}(g \circ f_T(a), g \circ f_\infty(a)) < \epsilon) \geq \mathbb{P}(\sup_{a\in\mathbb{A}} \delta_Y(f_T(a), f_\infty(a)) < \epsilon')$$

because the second implies the first $\forall \omega \in \Omega$. Hence, since pointwise convergence in probability $\lim_{T\to\infty} \mathbb{P}(\sup_{a\in\mathbb{A}} \delta_Y(f_T(a), f_\infty(a)) < \epsilon') = 1\ \forall \epsilon' > 0$ holds by assumption, it follows that

$$\lim_{T\to\infty} \mathbb{P}(\sup_{a\in\mathbb{A}} \delta_Z(g \circ f_T(a), g \circ f_\infty(a)) < \epsilon) = 1\ \forall \epsilon > 0.$$

Convergence a.s. follows since $\forall T \in \mathbb{N}$ it holds true, by (A.2) and Lemma A.56 that for every $\omega \in \Omega$, $\lim_{T\to\infty} \delta_Y(f_T(\omega, a), f_\infty(a)) < \epsilon'$ implies $\lim_{T\to\infty} \delta_Z(g \circ f_T(\omega, a), g \circ f_\infty(\omega, a)) < \epsilon$, and hence that,

$$\mathbb{P}(\lim_{T\to\infty} \sup_{a\in\mathbb{A}} \delta_{\mathbb{C}}(g \circ f_T(a), g \circ f_\infty(a)) < \epsilon) \geq \mathbb{P}(\lim_{T\to\infty} \sup_{a\in\mathbb{A}} \delta_Y(f_T(a), f_\infty(a)) < \epsilon')$$

because the second implies the first $\forall \omega \in \Omega$. Hence, since pointwise a.s. convergence holds by assumption, i.e. $\mathbb{P}(\lim_{T\to\infty} \sup_{a\in\mathbb{A}} \delta_Y(f_T(a), f_\infty(a)) < \epsilon') = 1 \, \forall \epsilon' > 0$, it follows that

$$\mathbb{P}(\lim_{T\to\infty} \sup_{a\in\mathbb{A}} \delta_Z(g \circ f_T(a), g \circ f_\infty(a)) < \epsilon) = 1.$$

□

Lemma A.58. (Heine-Cantor Theorem) [Davidson 1994, Theorem 2.19, p.28] *Let $(\mathbb{A}, \delta_{\mathbb{A}})$ and $(\mathbb{B}, \delta_{\mathbb{B}})$ be metric spaces and $f : \mathbb{A} \to \mathbb{B}$ be a continuous map at every $a \in \mathbb{A}$. Then, if $\mathbb{A}$ is compact, f is uniformly continuous on $\mathbb{A}$.*

Lemma A.59. (Convergence of Sieve Estimators) [Chen 2007, Theorem 3.1 and Remark 3.2 and White and Wooldrige 1991, Proposition 2.4 and Corollary 2.6] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}, \delta_{\mathbb{A}})$ be a metric space with Borel σ-algebra $\mathfrak{B}(\mathbb{A})$ and $\{\mathbb{A}_T\}_{T\in\mathbb{N}}$ be a sequence of compact subsets of $\mathbb{A}$ such that $\mathrm{cl}\Big(\bigcup_{T\in\mathbb{N}} \mathbb{A}_T\Big) \supseteq \mathbb{A}$. Suppose that the sequence $\{f_T\}_{T\in\mathbb{N}}$ of functions $f : \Omega \times \mathbb{A} \to \mathbb{R}$, continuous of $\mathbb{A} \, \forall T \in \mathbb{N}$ and such that, $\lim_{T\to\infty} \mathbb{P}\Big(\sup_{a\in\mathbb{A}} |f_T(a) - f(a)| > \epsilon\Big) = 0 \; \forall \epsilon > 0$, for some continuous deterministic function $f : \mathbb{A} \to \mathbb{R}$ satisfying, $f(a_0) = 0$ and $\inf_{a\in S^c_{a_0}(\epsilon)} |f(a_0) - f(a)| > 0 \; \forall \epsilon > 0$. Let $\hat{a}_T : \Omega \to \mathbb{A}$ be an $\mathcal{F}/\mathfrak{B}(\mathbb{A})$-measurable map such that, $f_T(\hat{a}_T) \leq \inf_{a\in\mathbb{A}_T} f_T(a) + O_p(\eta_T)$ with $\eta_T \to 0$ as $T \to \infty$. Then, $\lim_{T\to\infty} \Big(\delta_{\mathbb{A}}(\hat{a}_T, a_0) > \epsilon\Big) = 0 \, \forall \epsilon > 0$.*

Lemma A.60. (Compactification) [Dudley 2002, Theorem 2.8.2, p.72] *Any separable metric space $(\mathbb{A}, \delta_{\mathbb{A}})$ has a totally bounded metrization, i.e. there exists a metric $\delta'_{\mathbb{A}}$ on $\mathbb{A}$ inducing the same topology as $\delta_{\mathbb{A}}$ on $\mathbb{A}$ such that $(\mathbb{A}, \delta'_{\mathbb{A}})$ is totally bounded, so that the completion for $\delta'_{\mathbb{A}}$ is a compact metric space and a compactification of $\mathbb{A}$.*

Lemma A.61. (Weak Convergence on Product Spaces) [van der Vaart and Wellner 1996, Theorem 1.4.8, p. 32] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}_i, \mathcal{T}_{\mathbb{A}_i})$ be a topological space for every i on a countable set $\mathbb{I}$ and $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ be the product space $\mathbb{A} = \times_{i\in\mathbb{I}} \mathbb{A}_i$ with product topology $\mathcal{T}_{\mathbb{A}_i}$. Let $\{X_T(\omega)\}_{T\in\mathbb{N}}$ with $X_T : \Omega \to \mathbb{A}$ be a sequence in $\mathbb{A}$ for every $\omega \in \Omega$ and $X : \Omega \to \mathbb{A}$ be a separable random element. Then X_T converges weakly to X if and only if $(X_{T,i_1}, ..., X_{T,i_n})$ converges weakly to $(X_{i_1}, ..., X_{i_n})$ for every $n \in \mathbb{N}$.*

Definition A.62. (Separable Process) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $\mathbb{A}$ be a separable set, $(\mathbb{B}, \mathfrak{B})$ be a measurable space, and $X : \Omega \times \mathbb{A} \to \mathbb{B}$ be an $\mathcal{F}/\mathfrak{B}$-measurable element (a stochastic process) for every $a \in \mathbb{A}$. Then X is said to be*

separable with respect to $\mathbb{A}'$ if $\mathbb{A}'$ is a countable dense subset of $\mathbb{A}$, and there is a measure-zero set $\Omega^* \subset \Omega$, $\mathbb{P}(\Omega^*) = 0$, such that for every $\omega \notin \Omega^*$, $X(\omega, \cdot)$ is almost surely $\mathbb{A}'$-separable.

Definition A.63. (Separable Map) *Let $(\mathbb{A}, \delta_{\mathbb{A}})$ and $(\mathbb{B}, \delta_{\mathbb{B}})$ be metric spaces, and $\mathbb{A}$ be separable. Let $\mathbb{A}'$ be a countable, dense subset of $\mathbb{A}$. A function $f : \mathbb{A} \to \mathbb{B}$ is $\mathbb{A}'$-separable, or separable with respect to $\mathbb{A}'$, if $\forall a \in \mathbb{A}$, there exists a sequence $a_i \in \mathbb{A}'$ such that $a_i \to \mathbb{A}$ and $f(a_i) \to f(a)$.*

Remark A.64. *We cannot easily guarantee that a process is separable. We can however "turn" a non-separable process, into a separable process with the same finite-dimensional distributions (Lemma A.65).*

Lemma A.65. (Separable Modification) [Gusak et al. 2010, Theorem 3.2, p.22] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}, \delta_{\mathbb{A}})$ be a separable metric space, $(\mathbb{B}, \delta_{\mathbb{B}})$ a compact metric space, and $X : \Omega \times \mathbb{A} \to \mathbb{B}$ a stochastic process. Then there exists a separable version $\tilde{X} : \Omega \times \mathbb{A} \to \mathbb{B}$ of X. This is called a separable modification of X.*

Definition A.66. (Stochastic Process Versioning) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $\mathbb{A}$ be a set and $(\mathbb{B}, \mathfrak{B})$ be a measurable space. Two stochastic processes $X : \Omega \times \mathbb{A} \to \mathbb{B}$ and $Y : \Omega \times \mathbb{A} \to \mathbb{B}$ are said to be versions of one another if $\forall a \in \mathbb{A}$, $\mathbb{P}(\omega : X(\omega, a) = Y(\omega, a)) = 1$.*

Remark A.67. *If stochastic processes X and Y are versions of one another, they have the same finite-dimensional distributions.*

Remark A.68. *In Lemma A.65, if $\mathbb{B}$ is not compact, there still exists a separable version of X in some compactification $\tilde{\mathbb{B}}$ of $\mathbb{B}$.*

Definition A.69. (Compact Topological Space) *A set $\mathbb{A}'$ in a topological space $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ is compact if every covering of $\mathbb{A}'$ by open sets contains a finite sub-cover. $\mathbb{A}$ is a compact space if it is itself a compact set.*

Definition A.70. (Covering Number) *Let $(\mathbb{A}, \delta_{\mathbb{A}})$ be a metric space. The ϵ-covering number of $\mathbb{A}' \subseteq \mathbb{A}$ is the smallest number of open balls of radius ϵ that cover $\mathbb{A}'$.*

Definition A.71. (Totally Bounded Metric Space) *A metric space $(\mathbb{A}, \delta_{\mathbb{A}})$ is called totally bounded if and only if for every $\epsilon > 0$ there is a finite set $\mathbb{A}_F \subseteq \mathbb{A}$ such that for every $a \in \mathbb{A}$ there exists some $a_F \in \mathbb{A}_F$ such that $\delta_{\mathbb{A}}(a, a_F) < \epsilon$.*

Definition A.72. (Tight Probability Measures and Maps) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ be a topological space with Borel σ-algebra $\mathfrak{B}_{\mathbb{A}}$ generated by $\mathcal{T}_{\mathbb{A}}$ and $\mathbb{P}_{\mathbb{A}}$ be the probability measure on $\mathfrak{B}_{\mathbb{A}}$ induced on $\mathbb{A}$ by the measurable*

map $X : \Omega \to \mathbb{A}$. The probability measure $\mathbb{P}_{\mathbb{A}}$ is said to be tight if for every $\epsilon > 0$ there exists a compact set $\mathbb{A}_0 \subseteq \mathbb{A}$ with $\mathbb{P}_{\mathbb{A}}(\mathbb{A}_0) =\geq 1 - \epsilon$. The measurable map X is called tight if $\mathbb{P}_{\mathbb{A}}(X) = \mathbb{P} \circ X^{-1}$ is tight.

Definition A.73. (Asymptotically Tight Sequences) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and $(\mathbb{A}, \delta_{\mathbb{A}})$ be a metric space with Borel σ-algebra $\mathfrak{B}_{\mathbb{A}}$. Then a sequence $\{X_n\}_{n\in\mathbb{N}}$ of measurable maps $X_n : \Omega \to \mathbb{A}\ \forall n \in \mathbb{N}$ is said to be asymptotically measurable if for every $\epsilon > 0$ there exists a compact set $\mathbb{A}_0 \subseteq \mathbb{A}$ such that $\liminf \mathbb{P}(X_n \in \mathbb{A}_0^{\delta}) \geq 1 - \epsilon$ for every $\delta > 0$ where $\mathbb{A}_0 := \{a \in \mathbb{A} : \delta_{\mathbb{A}}(a, \mathbb{A}) < \delta\}$.*

Lemma A.74. (Asymptotically Tight Sequence) [van der Vaart and Wellner 1996, Lemma 1.3.8, p.21] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ be a topological space with Borel σ-algebra $\mathfrak{B}_{\mathbb{A}}$ and $\{X_n\}_{n\in\mathbb{N}}$ denote a sequence of measurable maps $X_n : \Omega \to \mathbb{A}\ \forall n \in \mathbb{N}$ satisfying $X_n \xrightarrow{d} X$ where $X : \Omega \to \mathbb{A}$ is a limit measurable map. Then $\{X_n\}_{n\in\mathbb{N}}$ is asymptotically tight if and only if X is tight.*

Lemma A.75. (Tightness and Separability) [van der Vaart and Wellner 1996, Lemma 1.3.2, p.17] *Let $(\mathbb{A}, \delta_{\mathbb{A}})$ be a metric space. A Borel probability measure on $(\mathbb{A}, \delta_{\mathbb{A}})$ is pre-tight if and only if it is separable. Let $(\mathbb{A}, \delta_{\mathbb{A}})$ be a complete metric space. For a Borel probability measure on $(\mathbb{A}, \delta_{\mathbb{A}})$, separability, pre-tightness and tightness are equivalent. Any Polish Borel probability measure is tight.*

Lemma A.76. (Tightness on Product Spaces) [van der Vaart and Wellner 1996, Lemma 1.4.3, p.30] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ be topological spaces with Borel σ-algebra $\mathfrak{B}_{\mathbb{A}}$ and $\mathfrak{B}_{\mathbb{B}}$ respectively. Let $\{X_n\}_{n\in\mathbb{N}}$ and $\{Y_n\}_{n\in\mathbb{N}}$ denote measurable sequences of maps $X_n : \Omega \to \mathbb{A}$ and $Y_n : \Omega \to \mathbb{B}\ \forall n \in \mathbb{N}$ respectively. Then $\{(X_n, Y_n)\}_{n\in\mathbb{N}}$ is asymptotically tight if and only if both $\{X_n\}_{n\in\mathbb{N}}$ and $\{Y_n\}_{n\in\mathbb{N}}$ are asymptotically tight.*

Lemma A.77. (Tight Gaussian Process) *If the gaussian sequence and mean zero and satisfies a uniform bound on the variance then it is tight.*

Proposition A.78. (Degenerate Weak Convergence Implies Convergence in Probability) *Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a complete probability space, $(\mathbb{A}, \delta_{\mathbb{A}})$ be a measurable metric space, $\{X_T\}_{T\in\mathbb{N}}$ be a sequence of $\mathbb{A}$-valued random variables $X_T : \Omega \to \mathbb{A}$, and $X : \Omega \to \mathbb{A}$ be some $\mathbb{A}$-valued random variable. Then $X_T \xrightarrow{p} X$ if $X_T \xrightarrow{d} X$ and X is degenerate.*

Proof. Immediate extension of the common result for real-valued random variables that can be found e.g. in Davidson (1994, Theorem 22.5, p.349) and Potscher and Prucha (2001, Theorem 10, p.209). □

Lemma A.79. (Arzelà-Ascoli Theorem) [Dudley 2002, Theorem 2.4.7, p.52] *Let* $(\mathbb{A}, \delta_{\mathbb{A}})$ *be a compact metric space,* $(\mathbb{C}(\mathbb{A}), \delta_{\mathbb{C}}^{\sup})$ *be the space of real-valued continuous functions defined on* $\mathbb{A}$ *with* sup*-norm. A subset* $\mathbb{C}' \subset \mathbb{C}(\mathbb{A})$ *is totally bounded if and only if it is uniformly bounded and equicontinuous, i.e. uniformly equicontinuous.*

Remark A.80. *The Arzelà-Ascoli Theorem above yields that a sequence* $\{f_n\}$ *of real-valued functions defined on a totally bounded metric space* $(\mathbb{A}, \delta_{\mathbb{A}})$ *satisfies the convergence* $\sup_{a\in\mathbb{A}} |f_n(a)| \to 0$ *if and only if* $f_n(a) \to 0 \, \forall a \in \mathbb{A}_0$ *where* $\mathbb{A}_0$ *is a dense subset of* $\mathbb{A}$ *and* $\{f_n\}$ *is asymptotically uniformly equicontinuous.*

Remark A.81. *Bounded linear operators and uniform Lipschitz classes constitute simple examples of families of functions satisfying the uniform equicontinuity condition in Lemma A.79 above.*

Definition A.82. (Purely Dimensional Sieves) *Let* $(\mathbb{A}, \delta_{\mathbb{A}})$ *be a metric space and* $\{\mathbb{A}_n\}_{n\in\mathbb{N}}$ *denote a sequence of compact subsets of* $\mathbb{A}$*. Furthermore, let* $\{f_n\}_{n\in\mathbb{N}}$ *denote a sequence of continuous maps* $f_n : \mathbb{A} \to \mathbb{R} \, \forall n \in \mathbb{N}$ *on* $\mathbb{A}$ *where* $\mathbb{R}$ *denotes the set of real numbers with its natural ordering. The sequence of subsets* $\{\mathbb{A}_n\}_{n\in\mathbb{N}}$ *of* $\mathbb{A}$ *is said to be purely dimensional w.r.t. the sequence of maps* $\{f_n\}_{n\in\mathbb{N}}$ *if and only if the sequence* $\{a_n\}_{n\in\mathbb{N}}$ *of minimizers* $a_n \in \arg\min_{a\in\mathbb{A}} f_n(a) \, \forall n \in \mathbb{N}$ *satisfies* $\pi_{\mathbb{A}_n}(a_n) \in \mathrm{int}_{\mathrm{lin}}(\mathbb{A}_n)$*, where* $\pi_{\mathbb{A}_n}(a_n)$ *denotes the metric projection of* a_n *on the subset* $\mathbb{A}$ *and* $\mathrm{int}_{\mathrm{lin}}$ *denotes the interior of* $\mathbb{A}_n$ *w.r.t. its linear span* $\mathrm{lin}(\mathbb{A}_n)$*.*[1]

Remark A.83. *Purely dimensional sieves are essentially sieves that impose only 'dimensionality restrictions' on the optimization problem. In other words, sieves do not have to 'grow in size', only in dimensions. The given name is thus justified by noting that, in applications, this property will most often be related to the formulation of sieves that increase in dimension but that do not constrain the optimization procedure within any of the given dimensions. Figure A.1 below plots for a fixed minimizer* $a \in \mathbb{A}$ *a very simple example (left) where increasing from one to two dimensions is sufficient, as well as, the opposite case (right) where the sieves need only to increase in size.*

Remark A.84. *In essence, the definition above requires the sieves* $\{\mathbb{A}_n\}_{n\in\mathbb{N}}$ *to be such that the global (unrestricted) minimizer of the function* f_n *to be always an element of the corresponding sieve* $\mathbb{A}_n$*, for every* $n \in \mathbb{N}$*. This is obviously not strictly related to any dimensionality issue as alternative sequences* $\{f_n\}_{n\in\mathbb{N}}$ *and* $\{\mathbb{A}_n\}_{n\in\mathbb{N}}$ *can be formulated so as to satisfy this condition without any mentioning of dimensions of* $\mathbb{A}$*.*

[1] By construction, the elements of the sequence of projection minimizers $\{\pi_{\mathbb{A}_n}(a_n)\}$ are not in the interior of $\mathbb{A}_n$ when the interior is defined w.r.t. the space $\mathbb{A}$.

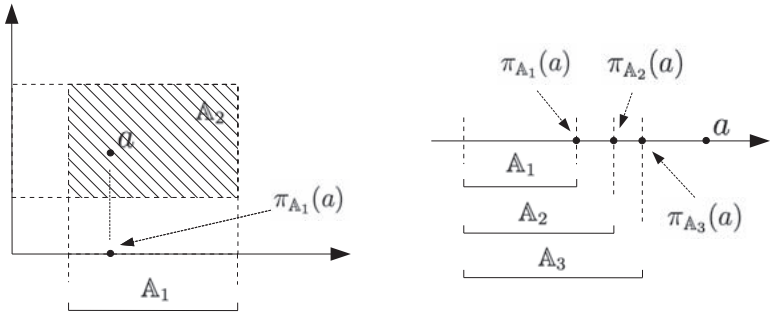

Figure A.1: Left: $\mathbb{A}_1$ is a subsection of the horizontal axis. $\mathbb{A}_2$ is the highlighted area. Pure dimensionality of sieves is reflected by the fact that $\pi_{\mathbb{A}_1}(a)$ is in the interior of $\mathbb{A}_1$ w.r.t. to the line. Right: $\mathbb{A}_1$, $\mathbb{A}_2$ and $\mathbb{A}_3$ are increasing subsets of the line. This time, projections are on the boundary of each sieve w.r.t. the line.

Lemma A.85. (Weierstrass's Extreme Value Theorem) [Munkres 2000, Theorem 27.4, p.174] *Let $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ be topological spaces with $\mathbb{Y}$ an ordered set in the order topology and $f : \mathbb{A} \to \mathbb{B}$ be a continuous map. If X is compact, then there exists points $(a', a'')in\mathbb{A} \times \mathbb{A}$ such that $f(a') \leq f(a) \leq f(a'')$ for every $a \in \mathbb{A}$.*

Definition A.86. (Schauder Basis) *Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ be a Banach space. A sequence $\{a_n\}_{n\in\mathbb{N}} \subset \mathbb{A}$ is a Schauder basis of $\mathbb{A}$ if for every $a \in \mathbb{A}$ there is a unique sequence of scalars $\{r_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ such that $a = \lim_{N\to\infty} \sum_n^N r_n a_n$.*

The following is thus an immediate corollary of Definition A.86

Corollary A.87. (Linear Independence of Schauder Basis) *Any finite collection of elements of the Schauder basis of a vector space consists of a set of linearly independent vectors.*

Remark A.88. *Examples of spaces with Schauder basis are: The standard bases of C^0 and L_p for $1 \leq p < \infty$ are Schauder bases. Every orthonormal basis in a separable Hilbert space is a Schauder basis. The Haar system is an example of a basis for $L_p(0,1)$ with $1 \leq p < \infty$. The Banach space $C([0,1])$ of continuous functions on the interval $[0,1]$, with the supremum norm, admits a Schauder basis. A Banach space with a Schauder basis is necessarily separable, but the converse is false. Every Banach space with a Schauder basis has the approximation property.*

Proposition A.89. (Geometric Sums of Random Variables) *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ and $(\mathbb{B}_i, \|\cdot\|_{\mathbb{B}_i})$ be complete normed vector spaces for every $i \in \mathbb{I}$ with $\mathbb{I}$ a countable set. Let $\{f_{i,n}\}_{i\in\mathbb{I}}$ be a sequence of measurable maps $f_{i,n} : \Omega \times \mathbb{A} \to \mathbb{B}_i$ for every $n \in \mathbb{N}$. Finally, suppose that $\sup_{i\in\mathbb{I}} \|f_{i,n}(a) - f_i(a)\|_{\mathbb{B}} \xrightarrow{p} 0$ as $n \to \infty$ for every $a \in \mathbb{A}$ and every $i \in \mathbb{I}$. Then, if $\exists\, K < \infty$ such that $\sup_{i\in\mathbb{I}} \|f_i(a)\|_{\mathbb{B}} \leq K(a)$ for every $a \in \mathbb{A}$ it follows that the geometric sum of $f_{i,n}(a)$ over i is bounded in probability as $n \to \infty$, for every $a \in \mathbb{A}$.*

Proof. For every $a \in \mathbb{A}$ it holds true that,

$$\begin{aligned}
\mathbb{P}\Big(\Big\|\sum_{i\in\mathbb{I}}\frac{1}{2^i}f_{i,n}(a)\Big\|_{\mathbb{B}} < K^*\Big) &= \mathbb{P}\Big(\Big\|\sum_{i\in\mathbb{I}}\frac{1}{2^i}(f_{i,n}(a)-f_i(a)+f_i(a))\Big\|_{\mathbb{B}} < K^*\Big) \\
&\geq \mathbb{P}\Big(\Big\|\sum_{i\in\mathbb{I}}\frac{1}{2^i}(f_{i,n}(a)-f_i(a))\Big\|_{\mathbb{B}} + \Big\|\sum_{i\in\mathbb{I}}\frac{1}{2^i}f_i(a)\Big\|_{\mathbb{B}} < K^*\Big) \\
&\geq \mathbb{P}\Big(\sum_{i\in\mathbb{I}}\frac{1}{2^i}\big\|(f_{i,n}(a)-f_i(a))\big\|_{\mathbb{B}} < \frac{K^*}{2}\Big) + \mathbb{P}\Big(\sum_{i\in\mathbb{I}}\frac{1}{2^i}\big\|f_i(a)\big\|_{\mathbb{B}} < \frac{K^*}{2}\Big) \\
&\geq \mathbb{P}\Big(\sum_{i\in\mathbb{I}}\frac{1}{2^i}\sup_{i\in\mathbb{I}}\big\|(f_{i,n}(a)-f_i(a))\big\|_{\mathbb{B}} < \frac{K^*}{2}\Big) \\
&\quad + \mathbb{P}\Big(\sum_{i\in\mathbb{I}}\frac{1}{2^i}\sup_{i\in\mathbb{I}}\big\|f_i(a)\big\|_{\mathbb{B}} < \frac{K^*}{2}\Big) \to 1.
\end{aligned}$$

where the final convergence to one follows by selecting $K^*/2 = K(a)$ for every $a \in \mathbb{A}$. □

Lemma A.90. (Weak Convergence) [van der Vaart and Wellner 1996, Theorem 3.3.1] *Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be normed vector spaces. Let $f_n : \mathbb{A} \to \mathbb{B}$ be a random map for every $n \in \mathbb{N}$ and $f : \mathbb{A} \to \mathbb{B}$ be deterministic. Let $\hat{a}_n$ be a random element of $\mathbb{A}$ satisfying $f_n(\hat{a}_n) = o_p(n^{-1/2})$ and $a_0 \in \mathbb{A}$ satisfy $f(a_0) = 0$, with $\hat{a}_n \xrightarrow{p} a_0$. Suppose that there exists a tight random element Z such that $\sqrt{n}(f_n - f)(a_0) \xrightarrow{d} Z$. Furthermore, let $\sqrt{n}\|(f_n - f)(\hat{a}_n) - (f_n - f)(a_0)\|_{\mathbb{B}} = o_p(1 + \sqrt{n}\|\hat{a}_n - a_0\|_{\mathbb{A}})$. Finally, suppose that f is Hadamard differentiable at a_0 with continuously invertible derivative $\nabla f(a_0, \cdot)$. Then $\sqrt{n}(\hat{a}_n - a_0) \xrightarrow{d} -\text{inv}\big(\nabla f(a_0, \cdot)\big)(Z)$.*

Lemma A.91. (Separable Metric Space and Unit Cube are Homeomorphic) [Davidson (1994, Theorem 6.22))] *Every separable metric space is homeomorphic to the unit cube with product topology.*

Lemma A.92. (Weierstrass Theorem) [Powell (1981, Theorem 6.1, p.61)] *For every function $f : \mathbb{R} \to \mathbb{R}$ such that $f \in \mathbb{C}[a, b]$ and every $\epsilon > 0$, there exists a polynomial of order k denoted $p_k \in \mathcal{P}_k[a, b]$, i.e. a function $p_k(x) = \sum_{i=1}^{k} \theta_i x^i$, such that $\|p_k - f\|_\infty < \epsilon$.*

Appendix B

Linear Operator Theory and Continuous Invertibility

Definition B.1. (Operator Norm) *Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *be normed vector spaces and* $\mathbb{L}(\mathbb{A}, \mathbb{B})$ *denote the space of bounded linear operators form* $\mathbb{A}$ *into* $\mathbb{B}$*. The operator norm* $\|\cdot\|_{\mathbb{L}_{\mathbb{A}}^{\mathbb{B}}}$ *on* $\mathbb{L}(\mathbb{A}, \mathbb{B})$ *is defined alternatively as*

$$\|f\|_{\mathbb{L}_{\mathbb{A}}^{\mathbb{B}}} := \sup_{a \in \mathbb{A}} \frac{\|f(a)\|_{\mathbb{B}}}{\|a\|_{\mathbb{A}}} \quad \text{or} \quad \|f\|_{\mathbb{L}_{\mathbb{A}}^{\mathbb{B}}} := \sup_{a \in \mathbb{A}: \|a\|_{\mathbb{A}} \leq 1} \|f(a)\|_{\mathbb{B}} \quad \text{for every} \quad f \in \mathbb{L}(\mathbb{A}, \mathbb{B}).$$

Lemma B.2. (Complete Dual Normed Vector Space) [Dudley 2002, Theorem 6.1.3, p.191] *For any normed vector space* $(\mathbb{A}, \|\cdot\|)$ *over* $\mathbb{B}$ *the dual space with operator norm* $(\mathbb{L}(\mathbb{A}, \mathbb{B}), \|\cdot\|_{\mathbb{L}_{\mathbb{A}}^{\mathbb{B}}})$ *is a Banach space.*

Lemma B.3. (Identification of $\mathbb{L}$) [Denkowski et al. 2003, Proposition 5.1.17, p.525] *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces and let the spaces* $\mathbb{L}(\mathbb{A}, \mathbb{B})$ *of bounded linear operators from* $\mathbb{A}$ *into* $\mathbb{B}$ *be equipped with the uniform norm. Then, the space* $\mathbb{L}(\mathbb{A}, \mathbb{L}(\mathbb{A}, \mathbb{B}))$ *is isometrically isomorphic to the space* $\mathbb{L}(\mathbb{A} \times \mathbb{A}, \mathbb{B})$ *of bounded bilinear operators from* $\mathbb{A} \times \mathbb{A}$ *into* $\mathbb{B}$ *with uniform norm.*

Lemma B.4. (Banach-Steinhaus Theorem) [Dudely, Theorm 6.5.1, p.212] *Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *be a Banach space and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *a normed vector space. Let* $f_T : \mathbb{A} \to \mathbb{B}$ *be a bounded linear operator for every* $T \in \mathbb{N}$*. If* $\sup_{T \in \mathbb{N}} \|f_T(a)\|_{\mathbb{B}} < \infty \; \forall a \in \mathbb{A}$ *then* $\sup_{T \in \mathbb{N}} \|f_T\| < \infty$ *in operator norm.*

Lemma B.5. (Linear Composition) [Winitzki (2010, Statement 2, p.28)] *Let* $\mathbb{A}$*,* $\mathbb{B}$ *and* $\mathbb{C}$ *be vector spaces and* $f : \mathbb{A} \to \mathbb{B}$ *and* $g : \mathbb{B} \to \mathbb{C}$ *be linear maps. Then the composition map* $h := g \circ f : \mathbb{A} \to \mathbb{C}$ *is linear.*

Lemma B.6. (Bounded Linear Operator) [Sviridyuk and Fedorov 2003, Theorem 1.1.1, p.3] *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces. Let an operator* $f : \mathbb{A} \to \mathbb{B}$ *be linear. Then the following statements are equivalent:* (i) *the operator*

f is continuous at one point; (ii) *the operator f is continuous;* (iii) *the operator f is bounded.*

Lemma B.7. (Linearity of the Pointwise Limit of a Sequence of Linear Functions) [Denkowski et al. 2003, Proposition 3.2.3, p.267] *If $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ is a Banach space, $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ is a normed vector space, $\{f_n\}_{n\in\mathbb{N}} \subseteq \mathbb{L}(\mathbb{A}, \mathbb{B})$ and for every $a \in \mathbb{A}$, $f(a) = \lim f_n(a)$ exists in $\mathbb{B}$, then $f \in \mathbb{L}(\mathbb{A}, \mathbb{B})$.*

Lemma B.8. (Injective Linear Operator) *Let $T : V \to W$ be a linear map. Then T is injective if and only if its kernel satisfies* $\mathrm{Ker}(T) = \{0\}$.

Lemma B.9. (Inverse of Linear Operator) [Kolmogorov and Fomin 1975, Theorem 1, p. 228, Luenberger 1997, Proposition 1, p.174] *The inverse of a linear operator between topological vector spaces is itself linear.*

Lemma B.10. (Bounded Inverse) [Kolmogorov and Fomin 1975, Theorem 2, p.229] *Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be Banach spaces and $f : \mathbb{A} \to \mathbb{B}$ be an invertible bounded linear operator $f \in \mathbb{L}(\mathbb{A}, \mathbb{B})$. Then the inverse operator $f^{-1} : \mathbb{B} \to \mathbb{A}$ is itself bounded.*

The following Corollary follows immediately from Lemmas B.10, B.6 and B.9.

Corollary B.11. (Continuous Inverse) *Let f be an invertible continuous linear operator. Then f^{-1} is continuous.*

Definition B.12. (Continuously Invertible Operator) *Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be normed vector spaces. A bounded linear operator $f \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ is said to be* continuously invertible *if its inverse is an operator $f^{-1} \in \mathbb{L}(\mathbb{B}, \mathbb{A})$, i.e. if it is a bounded linear operator defined on the range of f.*

Lemma B.13. (Continuous Invertibility and Bounded Inverse) [Sviridyuk and Fedorov 2003] *Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be Banach spaces and $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ be a bounded linear operator from $\mathbb{A}_f$ into $\mathbb{B}$, i.e. $f \in \mathbb{L}(\mathbb{A}_f, \mathbb{B})$. Then, the inverse operator $f^{-1} : f(\mathbb{A}_f) \to \mathbb{A}_f$ exists and is bounded on $f(\mathbb{A}_f)$ if and only if there exists $m \in \mathbb{R}_+$ such that $\|f(a)\|_{\mathbb{B}} \geq m\|a\|_{\mathbb{A}} \ \forall a \in \mathbb{A}_f$.*

Definition B.14. (Uniformly Continuously Invertible Operator) *Let $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ and $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ be Banach spaces and $f_i \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ for every $i \in \mathbb{I}$. Building on Lemma B.13, the family $f : \mathbb{I} \times \mathbb{A} \to \mathbb{B}$ is said to be continuously invertible uniformly in $i \in \mathbb{I}$ if and only if $f_i : \mathbb{L}(\mathbb{A}, \mathbb{B})$ is continuously invertible for every $i \in \mathbb{I}$ and there exists $m \in \mathbb{R}_+$ such that $\|f_i(a)\|_{\mathbb{B}} \geq m\|a\|_{\mathbb{A}} \ \forall (a, i) \in \mathbb{A} \times \mathbb{I}$.*

Proposition B.15. (Continuous Invertibility with Product Topology) *Let $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ be topological vector spaces for every i in some countable index set $\mathbb{I}$.*

Furthermore, let $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *denote the product space* $\mathbb{B} := \times_{i\in\mathbb{I}}\mathbb{B}_i$ *with product topology* $\mathcal{T}_{\mathbb{B}}$*. Then, a bounded linear operator* $f \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ *is continuously invertible if its projections* $f_i \in \mathbb{L}(\mathbb{A}, \mathbb{B}_i)$ *are continuously invertible for every* $i \in \mathbb{I}$*.*

Proof. Since every f_i is invertible, by Lemma A.44, every f_i is also bijective on its range. By the same argument as in Lemma A.46 and Proposition A.48 f is also bijective on its range. By Lemma A.44 it is invertible. Finally, continuity of $f^{-1} \in \mathbb{L}(\mathbb{B}, \mathbb{A})$ follows under the product topology from Lemma A.47 and a recollection that continuity is implied by an open inverse. □

Proposition B.16. (Continuously Invertible Composition)) *Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$*,* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *and* $(\mathbb{C}, \|\cdot\|_{\mathbb{C}})$ *be Banach spaces and* $f \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ *and* $g \in \mathbb{L}(\mathbb{B}, \mathbb{C})$ *be continuously invertible bounded linear operators. Then* $h := g \circ f : \mathbb{A} \to \mathbb{C}$ *is also a continuously invertible bounded linear operator.*

Proof. By Lemma B.5, $h := g \circ f$ is a linear map. By Lemma B.13, $\exists (m_f, m_g) \in \mathbb{R}_+ \times \mathbb{R}_+$ such that $\|f(a)\|_{\mathbb{B}} \geq m_f\|a\|_{\mathbb{A}} \ \forall a \in \mathbb{A}$ and $\|g(b)\|_{\mathbb{C}} \geq m_g\|b\|_{\mathbb{B}} \ \forall b \in \mathbb{B}$. It thus follows that $\|h(a)\|_{\mathbb{C}} = \|g(f(a))\|_{\mathbb{C}} \geq m_g\|f(a)\|_{\mathbb{B}} \geq m_g m_f\|a\|_{\mathbb{A}} = m_h\|a\|_{\mathbb{A}}$ $\forall a \in \mathbb{A}$ and again by Lemma B.13 that h is continuously invertible. □

Proposition B.17. (Uniformly Continuously Invertible Composition)) *Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$*,* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *and* $(\mathbb{C}, \|\cdot\|_{\mathbb{C}})$ *be Banach spaces. Furthermore, let* $f_i \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ *be continuously invertible uniformly in* $i \in \mathbb{I}$ *and* $g \in \mathbb{L}(\mathbb{B}, \mathbb{C})$ *be continuously invertible. Then* $h_i := g \circ f_i : \mathbb{A} \to \mathbb{C}$ *is a bounded linear operator continuously invertible in* $i \in \mathbb{I}$*.*

Proof. By Lemma B.5, $h_i := g \circ f_i$ is a linear map for every $i \in \mathbb{I}$. By Lemma B.13 and Definition B.14, $\exists (m_f, m_g) \in \mathbb{R}_+ \times \mathbb{R}_+$ such that $\|f_i(a)\|_{\mathbb{B}} \geq m_f\|a\|_{\mathbb{A}}$ $\forall (a, i) \in \mathbb{A} \times \mathbb{I}$ and $\|g(b)\|_{\mathbb{C}} \geq m_g\|b\|_{\mathbb{B}} \ \forall b \in \mathbb{B}$. It thus follows that $\|h_i(a)\|_{\mathbb{C}} = \|g(f_i(a))\|_{\mathbb{C}} \geq m_g\|f_i(a)\|_{\mathbb{B}} \geq m_g m_f\|a\|_{\mathbb{A}} = m_h\|a\|_{\mathbb{A}} \ \forall (a, i) \in \mathbb{A} \times \mathbb{I}$ and again by Lemma B.13 that h_i is continuously invertible uniformly in $i \in \mathbb{I}$. □

Proposition B.18. *(Uniform Bound on Bounded Linear Operators) Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *be normed vector spaces, let* $f_T : \mathbb{A} \to \mathbb{R}$ *be a continuous linear map for every* $T \in \mathbb{N}$*, and let* $f : \mathbb{A} \to \mathbb{B}$ *be also a continuous linear map satisfying* $\|f(a)\| \geq c\|a\|_{\mathbb{A}} \forall a \in \mathbb{A}$*. Finally, suppose that* $f_T(a_T) \to f(a)$ *for every sequence* $a_T \to a \in \mathbb{A}$*. Then* $\exists\, c^* > 0$ *and* $T^* \in \mathbb{N}$ *such that* $\|f_T(a_T)\|_{\mathbb{B}} \geq c^*\|a_T\|_{\mathbb{A}}$ *for every* $T > T^*$ *and every sequence* $a_T \to a \in \mathbb{A}$*.*

Proof. The proof is immediate by noting that for large enough T^*, the elements of the sequence $\{f_T\}$ must satisfy e.g. $\|f_T(a_T)\|_{\mathbb{B}} \geq c/2\|a_T\|_{\mathbb{A}}$ for every $T > T^*$ and $a_T \to a \in \mathbb{A}$. Suppose by contradiction that $\forall T^* \in \mathbb{N}$, $\exists \{a_T\}$ satisfying $a_T \to a \in \mathbb{A}$

and some $T > T^*$ such that $\|f_T(a_T)\|_{\mathbb{B}} < c/2\|a_T\|_{\mathbb{A}}$. Note that by the reverse triangle inequality,

$$\|f_T(a_T) - f(a)\|_{\mathbb{B}} \geq \Big|\|f_T(a_T)\|_{\mathbb{B}} - \|f(a)\|_{\mathbb{B}}\Big|$$

holds for every $T \in \mathbb{N}$ and $a_T \to a$. Then, together with $f(a) \geq c\|a\|_{\mathbb{A}} \forall a \in \mathbb{A}$, this implies that, $\forall T^* \in \mathbb{N}$, $\exists \{a_T\}$ satisfying $a_T \to a \in \mathbb{A}$ and some $T > T^*$ such that

$$\|f_T(a_T) - f(a)\|_{\mathbb{B}} \geq \Big|\|f_T(a_T)\|_{\mathbb{B}} - \|f(a)\|_{\mathbb{B}}\Big| \geq \Big|c/2\|a_T\|_{\mathbb{A}} - c\|a\|_{\mathbb{A}}\Big|.$$

However, $\Big|c/2\|a_T\|_{\mathbb{A}} - c\|a\|_{\mathbb{A}}\Big| \to c/2\|a\|_{\mathbb{A}}$ contradicts the assumption that $\|f_T(a_T) - f(a)\|_{\mathbb{B}} \to 0$ for every $a_T \to a$. □

Appendix C

Differentiability Concepts and Propositions

Typical concepts of differentiability are difficult to apply in the context of SNPII estimation with infinitely many auxiliary estimators. First, some smoothness conditions used in the convergence theorem are hard to obtain using differentiability concepts weaker than Fréchet differentiability (such as Gateaux or Hadamard differentiability). Second, Fréchet differentiability can not really be used without running into problems involving the lack on norms on some spaces. In particular, the concept of Fréchet differentiability does not apply to the infinite vector of auxiliary statistics because the auxiliary space $\mathcal{B}$ is not normed. Third, we will often need to 'extend' these traditional notions of differentiability to hold over sequences of functions, differentiability points and directions.

A satisfactory answer to all of these problems seems to consist of *(i)* making use of Hadamard differentiability, *(ii)* adding sufficient conditions for Fréchet differentiability to hold for some operators between normed spaces, and *(iii)* introducing a number of novel smoothness concepts that 'extend' Hadamard differentiability to hold over appropriate sequences.

C.1 Notation

In infinite dimensional spaces, care is needed in the definition of *derivative function* and *partial derivative function*. The same applies to *second-order differentiability* and *second-order partial differentiability*. Indeed, since alternative definitions exist in the literature (see Ren and Sen (2001, Remark 1) and references therein) it is often hard to understand what is meant by any of these concepts outside a specific context. To avoid ambiguity, the notational convention established below and in Definition C.3 applies throughout.

Unless explicitly stated otherwise, given a pair of topological vector spaces $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ and $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ and a map $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ the map $\nabla_{\mathbb{A}_0} f : \mathbb{A}_\nabla \to \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ denotes the *derivative function* of f tangentially to $\mathbb{A}_0$. The precise notion of derivative may vary according to Definition C.3. Typically, the derivative function $\nabla_{\mathbb{A}_0} f$ is a map from the set $\mathbb{A}_\nabla \subseteq \mathbb{A}_f$ of points at which f is differentiable into the space of bounded linear functionals $\mathbb{L}(\mathbb{A}_0, \mathbb{B})$.

The map $\nabla_{\mathbb{A}_0} f(a_\nabla) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ is called the *derivative* of f at a_∇. It is often (see Definition C.3) a bounded linear operator, i.e. an element of $\mathbb{L}(\mathbb{A}_0, \mathbb{B})$, for any $a_\nabla \in \mathbb{A}_\nabla$. The *directional derivative* $\nabla_{\mathbb{A}_0} f(a_\nabla, a_0)$ is thus a point in $\mathbb{B}$ for every $a_0 \in \mathbb{A}_0$. For completeness, $\nabla_{\mathbb{A}_0} f(a_\nabla, a_0)$ can be called the *derivative of f at a_∇ in the direction of a_0*.

For conciseness, $\nabla_{\mathbb{A}_0} f(a_\nabla)$ might also be denoted $\nabla_{\mathbb{A}_0} f_{a_\nabla}$. In this case, the *directional derivative* can be denoted $\nabla_{\mathbb{A}_0} f_{a_\nabla}(a_0)$. When the tangent set is not relevant, we might make use of the more condensed notations $\nabla_{a_0} f_{a_\nabla}$, sometimes $f^{\nabla_{a_0}}(a_\nabla)$ or even $f_{a_\nabla}^{\nabla_{a_0}}$. If the tangent set $\mathbb{A}_0$ coincides with the entire set $\mathbb{A}$ or its linear span $\mathrm{lin}(\mathbb{A})$, then it is omitted from the notation. Thus $\nabla f(a_\nabla) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ denotes the derivative of f at a_∇ tangentially to $\mathrm{lin}(\mathbb{A})$.

Remark C.1. *The ability to use concise notation when differentiability points or directions are not of interest is important to improve readability and keep formulas at an acceptable size. In any case, the reader shall be often reminded of these details when notation changes throughout the text.*

The map $\nabla_{\mathbb{A}_0} f(\cdot, a_0)$ is called a *directional derivative function* for every fixed $a_0 \in \mathbb{A}_0$. It is also called a *partial derivative function* whenever a_0 is an element of the vector of basis vectors $\mathbb{S}_{\mathbb{A}}$ that span $\mathbb{A}$. Note also, that we shall often work with the *vector of partial derivative functions*, denoted $\nabla_{\mathbb{A}_0} f(\cdot, \mathbb{S}_{\mathbb{A}})$. For conciseness, the notation $\nabla_{\mathbb{S}_{\mathbb{A}}} f$ or $f^{\nabla_{\mathbb{S}_{\mathbb{A}}}}$ is also adopted.

Remark C.2. *An important notational exception consists of $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta})$ which shall always denote an infinite vector, despite the finiteness of the number of basis vectors $\mathbb{S}_{\Theta_T}$. In particular, $\nabla_{\mathbb{S}_{\Theta_T}} Q_T(\boldsymbol{\theta})$ is a vector of partial derivatives at $\boldsymbol{\theta}$ in the directions $\nabla_{\mathbb{S}_{\Theta_T}}$ and a vector of zeros after the n^{th} entry, for every $n > |\mathbb{S}_{\Theta_T}|$.*

The notion of *second-order differentiability* of f is one requiring the differentiability of both $f : \mathbb{A}_f \to \mathbb{B}$ and $\nabla_{\mathbb{A}_0} f : \mathbb{A}_\nabla \to \mathbb{L}(\mathbb{A}_0, \mathbb{B})$. One should thus note that while $\nabla^2_{\mathbb{A}_0} f(a_\nabla) \in \mathbb{L}^2(\mathbb{A}_0 \times \mathbb{A}_0, \mathbb{B})$ we have that $\nabla^2_{\mathbb{A}_0} f(a_\nabla, a_0) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ denotes a linear operator, an element of $\mathbb{L}(\mathbb{A}_0, \mathbb{B})$, for every $(a_\nabla, a_0) \in \mathbb{A}_\nabla \times \mathbb{A}_0$.

Second-order derivative functions $\nabla^2_{\mathbb{A}_0} f : \mathbb{A}_\nabla \to \mathbb{L}^2(\mathbb{A}_0 \times \mathbb{A}_0, \mathbb{B})$ are always understood as maps from the set $\mathbb{A}_\nabla$ of points at which f is differentiable to the space

of bounded bilinear operators $\mathbb{L}^2(\mathbb{A}_0 \times \mathbb{A}_0, \mathbb{B})$.[1]

For the sake of intuition and clarity, just as in applications dealing with finite dimensional parameter spaces, we shall often require differentiability of *partial derivative functions* $\nabla_{a_0} f$, $a_0 \in \mathbb{S}_{\mathbb{A}}$ and hence deal with notions of *second-order partial differentiability.*

C.2 Some General Definitions and Results

Definition C.3. (Differentiability of Operators) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be metrizable topological vector spaces. Let* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *and consider for some* $(a_\nabla, a_0) \in \mathbb{A}_f \times \mathbb{A}$ *the limit of the sequence in* $\mathbb{B}$*,*

$$\nabla f(a_\nabla, a_0) := \lim_{t \to 0} \frac{f(a_\nabla + t a_0) + f(a_\nabla)}{t}. \tag{C.1}$$

If this limit exists, then it is called the first variation of f *at* $a_\nabla \in \mathbb{A}_f$ *in the direction of* $a_0 \in \mathbb{A}$*. If (C.1) holds for every direction* $a_0 \in \mathbb{A}$ *and* $\nabla f(a_\nabla, \cdot)$ *is linear, then* $f : \mathbb{A}_f \to \mathbb{B}$ *is said to be Gateaux differentiable at* a_∇*. Linearity is sometimes not required.*[2] *If* $f : \mathbb{A}_f \to \mathbb{B}$ *is Gateaux differentiable then* $\nabla f(a_\nabla, a_0)$ *is called the Gateaux derivative of* $f : \mathbb{A}_f \to \mathbb{B}$ *at* $a_\nabla \in \mathbb{A}_f$ *in the direction of* $a_0 \in \mathbb{A}$*.*

If the limit,

$$\nabla f(a_\nabla, a_0) := \lim_{t_n \to 0} \frac{f(a_\nabla + t_n a_n) + f(a_\nabla)}{t_n}$$

exists for every sequence $t_n \to 0$ *and* $a_n \to a_0$ *with* $a_\nabla + t a_n \in \mathbb{A}_f \, \forall n \in \mathbb{N}$ *and* $\nabla f(a_\nabla, \cdot) \in \mathbb{L}(\mathbb{A}, \mathbb{B})$*, then* $f : \mathbb{A}_f \to \mathbb{B}$ *is said to be Hadamard differentiable at* a_∇*. In this case* $\nabla f(a_\nabla, a_0) \in \mathbb{B}$ *is called the Hadamard derivative of* f *at* a_∇ *in the direction of* a_0*. This is equal to the Gateaux derivative.*[3]

On a normed vector space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *additional definitions of differentiability are available. In particular, a map* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *is said to be compact (bounded) differentiable if there exists a linear map* $\nabla f(a_\nabla, \cdot) : \mathbb{A} \to \mathbb{B}$ *such that,*

$$\lim_{t \to 0} \sup_{a_0 \in \mathbb{A}_0, a_\nabla + t a_0 \in \mathbb{A}_f} \left\| \frac{f(a_\nabla + t a_0) + f(a_\nabla)}{t} - \nabla f(a_\nabla, a_0) \right\|_{\mathbb{B}} = 0$$

holds for every compact (bounded) $\mathbb{A}_0 \subseteq \mathbb{A}$*.*[4]

[1] The more immediate definition is $\nabla^2_{\mathbb{A}_0} f : \mathbb{A}_\nabla \to \mathbb{L}(\mathbb{A}_0, \mathbb{L}(\mathbb{A}_0, \mathbb{B}))$. However, following Denkowski et al. (2003, Proposition 5.1.17, p.525) we note that $\mathbb{L}(\mathbb{A}_0, \mathbb{L}(\mathbb{A}_0, \mathbb{B}))$ can be identified with $\mathbb{L}^2(\mathbb{A}_0 \times \mathbb{A}_0, \mathbb{B})$.

[2] While homogeneity of first-degree holds by construction, linearity must be additionally assumed. Here we define differentiability always w.r.t. a linear map. Continuity is however not assumed and may fail in infinite dimensional spaces.

[3] Some authors do not require continuity of $f(a_\nabla, \cdot)$ in the definition of Hadamard differentiability.

[4] Again, some authors require $\nabla f(a_\nabla, \cdot)$ to be only homogeneous.

Finally, when both $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *are normed vector spaces, then a function* $f : \mathbb{A}_f \to \mathbb{B}$ *is said to be Fréchet differentiable at* a_∇ *if there exists a continuous linear functional* $\nabla f(a_\nabla, \cdot)$ *satisfying,*

$$\lim_{\|a_0\|_{\mathbb{A}} \to 0} \frac{\|f(a_\nabla + a_0) - f(a_\nabla) - \nabla f(a_\nabla, a_0)\|_{\mathbb{B}}}{\|a_0\|_{\mathbb{A}}} = 0.$$

In this case, $\nabla f(a_\nabla, a_0)$ *is called the Fréchet derivative of* $f : \mathbb{A}_f \to \mathbb{B}$ *at* $a_\nabla \in \mathbb{A}_f$ *in the direction of* $a_0 \in \mathbb{A}$*. The Fréchet derivative function* $\nabla f(a_\nabla, \cdot)$ *is always continuous. This is not an assumption.*

Remark C.4. *When* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *is a metrizable topological space and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *is a normed vector space, then Hadamard differentiability is equivalent to compact differentiability with continuous derivative. Also, in this case, a Hadamard differentiable function satisfies* $\|f(a_\nabla + t_n a_n) - f(a_\nabla) - \nabla f(a, a_0)\| = o(t_n)$ *for every* $t_n \to 0$ *and* $a_n \to a_0$ *with* $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$*. When both* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *are normed vector spaces, then Fréchet differentiability is equivalent to bounded differentiability with continuous derivative.*

Alternative definitions of *compact* and *bounded* differentiability that do not require the image space of the smooth operator to be normed are also available. These definitions also satisfy the equivalence relations stated in Remark C.4.

Definition C.5. (Differentiability of Operators) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be metrizable topological vector spaces. Then, the map* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *is said to be compact (bounded) differentiable if there exists a linear map* $\nabla f(a_\nabla, \cdot) : \mathbb{A} \to \mathbb{B}$ *such that,*

$$\lim_{t \to 0} \frac{f(a_\nabla + t a_0) + f(a_\nabla) - \nabla f(a_\nabla, t a_0)}{t} = 0$$

uniformly for a_0 *on compact (bounded) subsets* $\mathbb{A}_0 \subseteq \mathbb{A}$*.*

In what follows, Corollary C.6 below states an immediate implication of the definition of Hadamard differentiability encountered above.

Corollary C.6. (Differentiability Sub-Tangentially) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces. Let* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$*. Then,* f *is also Hadamard differentiable at* a_∇ *tangentially to* $\mathbb{A}_1 \subseteq \mathbb{A}_0$*.*

Remark C.7. *In the context of the above corollary, it is quite trivial to show that* $\nabla_{\mathbb{A}_1} f(a_\nabla, \cdot)$ *is simply a restriction of* $\nabla_{\mathbb{A}_0} f(a_\nabla, \cdot)$ *onto* $\mathbb{A}_1$ *and hence that* $\nabla_{\mathbb{A}_1} f(a, a') = \nabla_{\mathbb{A}_0} f(a, a') \; \forall (a, a') \in \mathbb{A} \times \mathbb{A}_1$*.*

Lemma C.8. (Derivative Uniqueness) [Luenberger 1997, Proposition 2, p. 173] *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces and* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at* $a_\nabla \in \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$. *Then there exists a unique continuous linear map* $\nabla_{\mathbb{A}_0} f(a_\nabla) : \mathbb{A}_0 \to \mathbb{B}$ *satisfying the definition of Hadamard derivative above.*

Lemma C.9. (Chain Rule) [van der Vaart and Wellner 1996, Lemma 3.9.3, p.373] *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$, $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *and* $(\mathbb{C}, \mathcal{T}_{\mathbb{C}})$ *be topological vector spaces. Let* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at every point of* $\mathbb{A}_\nabla \subseteq \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0$ *and let* $g : \mathbb{B}_g \subseteq \mathbb{B} \to \mathbb{C}$ *be Hadamard differentiable at* $\mathbb{B}_\nabla := f(\mathbb{A}_\nabla) \subseteq \mathbb{B}_g$ *tangentially to* $\mathbb{B}_0 := \nabla_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0)$. *Then,* $g \circ f : \mathbb{A}_f \to \mathbb{C}$ *is differentiable at every point of* $\mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$ *with derivative* $\nabla_{\mathbb{B}_0} g\Big(f(a_\nabla), \nabla_{\mathbb{A}_0} f(a_\nabla)\Big)$.

Lemma C.2.1. (Delta Method) [van der Vaart and Wellner 1996, Theorem 3.9.4, p.374] *Let* $\mathbb{A}$ *and* $\mathbb{B}$ *be metrizable topological vector spaces. Let* $f : \mathbb{A}_f \subset \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at* a_0 *tangentially to* $\mathbb{A}_0$. *Let* $X_n : \Omega_n \to \mathbb{A}_f$ *be maps with* $r_n(X_n - a_0) \xrightarrow{d} X$ *for some constants* $r_n \to \infty$, *where* X *is separable and takes its values in* $\mathbb{A}_0$. *Then* $r_n(f(X_n) - f(a_0)) \xrightarrow{d} f'_{a_0}(X)$. *If* f'_{a_0} *is defined and continuous on the whole of* $\mathbb{A}$ *then the sequence* $r_n(f(X_n) - f(a_0)) - f'_\theta(r_n(X_n - a_0))$ *converges to zero in outer probability.*

In extremum estimator theory dealing with smooth criterion functions, it is important to note that a Z-estimator formulation is also available. The following lemma provides the desired result.

Lemma C.10. (A Generalization of Fermat's Stationary Points Theorem) [Luenberger (1997)] *Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *and* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *be normed vector spaces and let* $f : \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at* $a_0 \in \text{int}(\mathbb{A})$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$, *with continuous linear Hadamard derivative denoted* $\nabla_{\mathbb{A}_0} f_{a_0} : \mathbb{A}_0 \to \mathbb{B}$. *Furthermore, suppose that* $\mathbb{B}$ *is a totally ordered set and let* a_0 *be a local minimizer of* f *on* $\mathbb{A}$, *i.e. there exists an open ball of radius* $\epsilon > 0$ *centered in* a_0, *denoted* $S_{a_0}(\epsilon) \subset \mathbb{A}$ *such that* $f(a_0) \leq f(a)\ \forall\, a \in S_{a_0}(\epsilon)$. *Then* $\nabla_{\mathbb{A}_0} f_{a_0}(a) = 0\ \forall\, a \in \mathbb{A}_0$. *If* $\nabla_{\mathbb{A}_0} f_{a_0}(a) \neq 0$ *for some* $a \in \mathbb{A}_0$ *then* a_0 *is not a local minimizer of* f. *Finally, if* $\mathbb{A}$ *has a basis* $\mathbb{S}_{\mathbb{A}}$, *then having* $\nabla_{\mathbb{A}_0} f_{a_0}(a) = 0\ \forall\, a \in \mathbb{S}_{\mathbb{A}}$ *ensures that* $\nabla_{\mathbb{A}_0} f_{a_0}(a) = 0\ \forall\, a \in \mathbb{A}_0$.

In the context of SNPII estimation with infinitely many auxiliary statistics we are often interested in deriving smoothness of an operator from the smoothness of its projections. The following proposition is thus important.

Proposition C.11. (Hadamard Differentiability with Product Topology) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ *be topological vector spaces for every* i *in some countable index set* $\mathbb{I}$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be the product space* $\mathbb{B} = \times_{i \in \mathbb{I}} \mathbb{B}_i$ *with product topology* $\mathcal{T}_{\mathbb{B}}$. *Then, a map*

$f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ is Hadamard differentiable at a point $a_\nabla \in \mathbb{A}_f$ tangentially to $\mathbb{A}_0 \subseteq \mathbb{A}$ if and only if the coordinate projection $\pi_i f : \mathbb{A}_f \to \mathbb{B}_i$ is also Hadamard differentiable at $a_\nabla \in \mathbb{A}_f$ tangentially to $\mathbb{A}_0 \in \mathbb{A}$ for every $i \in \mathbb{I}$.

Proof. By definition, f is Hadamard differentiable at $a_\nabla \in \mathbb{A}_f$ tangentially to $\mathbb{A}_0 \subseteq \mathbb{A}$ if and only if there exists a continuous linear functional $\nabla_{\mathbb{A}_0} f(a_\nabla) : \mathbb{A}_0 \to \mathbb{B}$ such that, every sequence $\left\{b_n(t_n, a_n)\right\}_{T \in \mathbb{N}} \subset \mathbb{B}$ defined as,

$$b_n(t_n, a_n) := \frac{f(a_\nabla + t_n a_n) - f(a_\nabla) - t_n \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)}{t_n}$$

converges to zero, for every $t_n \to 0$ and every $a_n \to a_0 \in \mathbb{A}_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \, \forall \, n \in \mathbb{N}$. Now, by Corollary A.16 given the product topology $\mathcal{T}_\mathbb{B}$ on $\mathbb{B}$, convergence of the sequence $b_n(t_n, a_n) \to 0$ on the product space $\mathbb{B}$ occurs if and only if its coordinate projections $\pi_i b_n(t_n, a_n)$ also converge $\pi_i b_n(t_n, a_n) \to 0$ in $\mathbb{B}_i$ for every $i \in \mathbb{I}$. By linearity and continuity of the coordinate projection (Lemma A.15 and Proposition A.50) and the definition of Hadamard derivative, it follows immediately that $\nabla \pi_i(\boldsymbol{\beta}) = \pi_i \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ for every $i \in \mathbb{I}$. By the chain rule we thus obtain,

$$\nabla_{\mathbb{A}_0} \pi_i f(a_\nabla, a_0) = \pi_i \nabla_{\mathbb{A}_0} f(a_\nabla, a_0).$$

As a result, we then have that $b_n(t_n, a_n) \to 0$ if and only if,

$$\begin{aligned} \pi_i b_n(t_n, a_n) : &= \pi_i \left(\frac{f(a_\nabla + t_n a_n) - f(a_\nabla) - t_n \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)}{t_n} \right) \\ &= \frac{\pi_i f(a_\nabla + t_n a_n) - \pi_i f(a_\nabla) - t_n \nabla_{\mathbb{A}_0} \pi_i f(a_\nabla, a_0)}{t_n} \to 0 \text{ for every } i \in \mathbb{I}. \end{aligned} \tag{C.2}$$

Finally, since by Lemma B.5 a composition of linear maps is linear, and by lemma A.29 a composition of continuous maps is continuous, $\pi_i \circ \nabla_{\mathbb{A}_0} f(a_\nabla)$ is a continuous linear map on $\mathbb{A}_0$. This implies, by definition, that the convergence in (C.2) above holds if and only if $\pi_i f$ is Hadamard differentiable for every $i \in \mathbb{I}$. Hence, the complete argument goes as follows: *(i)* f is Hadamard at a_∇ if and only if every sequence $b_n(t_n, a_n) \to 0$; *(ii)* every sequence $b_n(t_n, a_n) \to 0$ if and only if every sequence $\pi_i b_n(t_n, a_n) \to 0 \, \forall \, i \in \mathbb{I}$, and; *(iii)* every $\pi_i b_n(t_n, a_n) \to 0 \, \forall \, i \in \mathbb{I}$ if and only if every $\pi_i f$ is Hadamard at a_∇. We thus conclude that f is Hadamard at a_∇ if and only if $\pi_i f$ is Hadamard at a_∇. □

The following corollary follows immediately by continuity of continuous compositions (Lemma A.29), Proposition C.11 above, and the fact that $\pi_i \nabla_{\mathbb{A}_0} f(a_\nabla, a_0) = \nabla_{\mathbb{A}_0} \pi_i f(a_\nabla, a_0)$.

Corollary C.12. (Continuous Differentiability with Product Topology) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ *be topological vector spaces for every* i *in some countable index set* $\mathbb{I}$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be the product space* $\mathbb{B} = \times_{i\in\mathbb{I}}\mathbb{B}_i$ *with product topology* $\mathcal{T}_{\mathbb{B}}$*. Then, a map* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *is continuously Hadamard at every point of* $\mathbb{A}_\nabla \subseteq \mathbb{A}_f \subseteq \mathbb{A}$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$ *if and only if the coordinate projection* $\pi_i f : \mathbb{A}_f \to \mathbb{B}_i$ *is also continuously Hadamard differentiable at every point of* $\mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$ *for every* $i \in \mathbb{I}$*.*

Proposition C.13. (Twice Differentiable Compositions) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$*,* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$*,* $(\mathbb{C}, \mathcal{T}_{\mathbb{C}})$*,* $\Big(\mathbb{L}(\mathbb{A}_0, \mathbb{B}), \mathcal{T}_{\mathbb{L}^{\mathbb{B}}_{\mathbb{A}_0}}\Big)$ *and* $\Big(\mathbb{L}(\mathbb{B}_0, \mathbb{C}), \mathcal{T}_{\mathbb{L}^{\mathbb{C}}_{\mathbb{B}_0}}\Big)$ *be topological vector spaces where* $\mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *and* $\mathbb{L}(\mathbb{B}_0, \mathbb{C})$ *denote the spaces of bounded linear operators from* $\mathbb{A}_0 \subseteq \mathbb{A}$ *into* $\mathbb{B}$ *and* $\mathbb{B}_0 \subseteq \mathbb{B}$ *into* $\mathbb{C}$ *respectively. Let* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at every point of* $\mathbb{A}_\nabla \subseteq \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0$*, with derivative at* $a_\nabla \in \mathbb{A}_\nabla$ *in the direction of* $a \in \mathbb{A}_0$ *denoted* $\nabla_{\mathbb{A}_0} f(a_\nabla, a)$*, and* $g : \mathbb{B}_g \subseteq \mathbb{B} \to \mathbb{C}$ *be Hadamard differentiable at every point of* $\mathbb{B}_\nabla := f(\mathbb{A}_\nabla)$ *tangentially to* $\mathbb{B}_0 := \nabla_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0)$*. Then, the composition map* $h := g \circ f : \mathbb{A}_f \to \mathbb{C}$ *is Hadamard differentiable at every point of* $\mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$*. If furthermore, the derivative function* $\nabla_{\mathbb{A}_0} f : \mathbb{A}_\nabla \to \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *is Hadamard differentiable at every point of its domain* $\mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$*, and the map* $\nabla_{\mathbb{B}_0} g : \mathbb{B}_\nabla \times \mathbb{L}_\nabla(\mathbb{A}_0, \mathbb{B}) \to \mathbb{L}(\mathbb{A}_0, \mathbb{C})$ *is Hadamard differentiable at every point of its domain* $\mathbb{B}_\nabla \times \mathbb{L}_\nabla(\mathbb{A}_0, \mathbb{B}) := f(\mathbb{A}_\nabla) \times \nabla_{\mathbb{A}_0} f(\mathbb{A}_\nabla) \subseteq \mathbb{B} \times \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *tangentially to* $\mathbb{B}_0 \times \mathbb{L}_0(\mathbb{A}_0, \mathbb{B}) := \nabla_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0) \times \nabla^2_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0) \subseteq \mathbb{B} \times \mathbb{L}(\mathbb{A}_0, \mathbb{B})$*, then the derivative function* $\nabla_{\mathbb{A}_0} h := \nabla_{\mathbb{B}_0} g(f, \nabla_{\mathbb{A}_0} f) : \mathbb{A}_f \to \mathbb{L}(\mathbb{A}_0, \mathbb{C})$ *is also Hadamard differentiable at every point of* $\mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$*.*

Proof. Differentiability of $h : \mathbb{A} \to \mathbb{C}$ on $\mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0$ follows immediately from the chain-rule Lemma C.9. For completeness, the proof goes as follows. For every $a_\nabla \in \mathbb{A}_\nabla$ and every sequence $t_n \to 0$ and $a_n \to a_0 \in \mathbb{A}_0$ as $n \to \infty$ with $(a_0 + a_n t_n) \in \mathbb{A}_f \; \forall n \in \mathbb{N}$ it holds true that,

$$\frac{h(a_\nabla + t_n a_n) - h(a_0)}{t_n} = \frac{g\Big(f(a_\nabla + t_n a_n)\Big) - g\Big(f(a_\nabla)\Big)}{t_n} = \frac{g\Big(b_\nabla + b_n t_n)\Big) - g\Big(b_\nabla\Big)}{t_n} \tag{C.3}$$

where

$$b_\nabla = f(a_\nabla) \;\text{ and }\; b_n := \frac{f(a_\nabla + t_n a_n) - f(a_\nabla)}{t_n} \to b_0 := \nabla_{\mathbb{A}_0} f(a_\nabla, a_0) \;\; \forall t_n \to 0 \tag{C.4}$$

and $a_n \to a_0 \in \mathbb{A}_0$, with the convergence $b_n \to b_0 \in \mathbb{B}_0$ being implied by differentiability of $f : \mathbb{A}_f \to \mathbb{B}$ at $a_\nabla \in \mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0$. As a result, by differentiability of g at $b_\nabla \in \mathbb{B}_\nabla$ tangentially to $\mathbb{B}_0$,

$$\frac{g\Big(b_\nabla + t_n b_n)\Big) - g\Big(b_\nabla\Big)}{t_n} \to \nabla_{\mathbb{B}_0} g(b_\nabla, b_0) = \nabla_{\mathbb{B}_0} g\Big(f(a_\nabla), \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)\Big) \;\; \forall t_n \to 0$$

and $b_n \to b_0 \in \mathbb{B}_0$. By (C.3) this implies the desired result that,

$$\frac{h(a_\nabla + t_n a_n) - h(a_\nabla)}{t_n} \to \nabla_{\mathbb{A}_0} h(a_\nabla, a_0) \quad \forall t_n \to 0 \text{ and } a_n \to a_0 \in \mathbb{A}_0.$$

Now, differentiability of the derivative function $\nabla_{\mathbb{A}_0} h := \nabla_{\mathbb{B}_0} g(f, \nabla_{\mathbb{A}_0} f) : \mathbb{A}_f \to \mathbb{L}(\mathbb{A}_0, \mathbb{C})$ at every point of $\mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0$ follows by a similar argument. In particular, differentiability of $\nabla_{\mathbb{A}_0} f : \mathbb{A}_\nabla \to \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ at every point of $\mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0$ implies convergence of the sequence $\{L_n\}_{n\in\mathbb{N}} \subset \mathbb{L}(\mathbb{A}_0, \mathbb{B})$, defined below, to a point $L_0 \in \mathbb{L}_0(\mathbb{A}_0, \mathbb{B}) := \nabla^2_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0) \subseteq \mathbb{L}(\mathbb{A}_0, \mathbb{B})$,

$$L_n := \frac{\nabla_{\mathbb{A}_0} f(a_\nabla + t_n a_n) - \nabla_{\mathbb{A}_0} f(a_\nabla)}{t_n} \to L_0 := \nabla^2_{\mathbb{A}_0} f(a_\nabla, a_0) \quad \forall\ t_n \to 0$$

and $a_n \to a_0 \in \mathbb{A}_0$. Furthermore, differentiability of $\nabla_{\mathbb{B}_0} g : \mathbb{B}_\nabla \times \mathbb{L}_\nabla(\mathbb{A}_0, \mathbb{B}) \to \mathbb{L}(\mathbb{A}_0, \mathbb{C})$ at every point of its domain $\mathbb{B}_\nabla \times \mathbb{L}_\nabla(\mathbb{A}_0, \mathbb{B}) := f(\mathbb{A}_\nabla) \times \nabla_{\mathbb{A}_0} f(\mathbb{A}_\nabla)$ tangentially to $\mathbb{B}_0 \times \mathbb{L}_0(\mathbb{A}_0, \mathbb{B}) := \nabla_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0) \times \nabla^2_{\mathbb{A}_0} f(\mathbb{A}_\nabla, \mathbb{A}_0)$ implies that,

$$\frac{\nabla_{\mathbb{B}_0} g\Big((b_\nabla, L_\nabla) + t_n(b_n, L_n)\Big) - \nabla_{\mathbb{B}_0} g\Big((b_\nabla, L_\nabla)\Big)}{t_n} \to \nabla^2_{\mathbb{B}_0} g\Big((b_\nabla, L_\nabla)\,,\,(b_0, L_0)\Big)$$

for every sequence $t_n \to 0$ and $(b_n, L_n) \to (b_0, L_0) \in \mathbb{B}_0 \times \mathbb{L}_0(\mathbb{A}_0, \mathbb{B})$ with $(b_\nabla, L_\nabla) + t_n(b_n, L_n) \in \mathbb{B}_\nabla \times \mathbb{L}_\nabla(\mathbb{A}_0, \mathbb{B})\ \forall n \in \mathbb{N}$. The desired result now follows by noting precisely that,

$$\begin{aligned}
&\frac{\nabla_{\mathbb{A}_0} h(a_\nabla + t_n a_n) - \nabla_{\mathbb{A}_0} h(a_\nabla, \cdot)}{t_n} \\
&\qquad = \frac{\nabla_{\mathbb{B}_0} g\Big(f(a_\nabla + t_n a_n), \nabla_{\mathbb{A}_0} f(a_\nabla + t_n a_n)\Big) - \nabla_{\mathbb{B}_0} g\Big(f(a_\nabla), \nabla_{\mathbb{A}_0} f(a_\nabla)\Big)}{t_n} \\
&\qquad\qquad = \frac{\nabla_{\mathbb{B}_0} g\Big((b_\nabla, L_\nabla) + t_n(b_n, L_n)\Big) - \nabla_{\mathbb{B}_0} g\Big((b_\nabla, L_\nabla)\Big)}{t_n},
\end{aligned}$$

so that $\nabla_{\mathbb{A}_0} h := \nabla_{\mathbb{B}_0} g(f, \nabla_{\mathbb{A}_0} f) : \mathbb{A}_f \to \mathbb{L}(\mathbb{A}_0, \mathbb{C})$ is differentiable with derivative,

$$\nabla^2_{\mathbb{A}_0} h = \nabla^2_{\mathbb{B}_0} g\Big(\Big(f, \nabla_{\mathbb{A}_0} f\Big)\,,\,\Big(\nabla_{\mathbb{A}_0} f, \nabla^2_{\mathbb{A}_0} f\Big)\Big) : \mathbb{A}_f \to \mathbb{L}^2(\mathbb{A}_0 \times \mathbb{A}_0, \mathbb{C}).$$

□

Finally, let us define two forms of differentiability that are important for the theory that follows in Section C.3. Both can be found in van der Vaart (1995).

Definition C.14. (Continuous Hadamard Differentiability) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *be topological vector spaces, let* $\mathbb{A}_f \subset \mathbb{A}$. *A function* $f : \mathbb{A}_f \to \mathbb{B}$ *is said to be continuously Hadamard differentiable on a convex set* $\mathbb{A}_\nabla \subset \mathbb{A}_f$ *contained in a neighborhood of* a_∇ *if and only if* f *is Hadamard differentiable at every* $a \in \mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$,

and its derivatives satisfy $\lim_{a\to a'} \nabla_{\mathbb{A}_0} f(a, a_0) = \nabla_{\mathbb{A}_0} f(a', a_0)$ *for every* $a_0 \in \mathbb{A}_0$ *for every* $a' \in \mathbb{A}_f$ *and* $\lim_{a'\to a_\nabla} \nabla_{\mathbb{A}_0} f(a', a_0) = \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)$ *uniformly in* a_0 *in totally bounded sets of* $\mathbb{A}_0$.[5]

Following van der Vaart (1995) we define uniform differentiability as follows.

Definition C.15. (Uniform Hadamard Differentiability) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *be topological vector spaces and* $\mathbb{A}_f \subset \mathbb{A}$ *be convex. Then* $f : \mathbb{A}_f \to \mathbb{B}$ *is said to be uniformly Hadamard differentiable along every sequence* $a_n \to a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0$ *if and only if,*

$$t_n^{-1}\Big(f(a_n + t_n a_n') - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)\Big) \to 0$$

holds for every sequence $t_n \to 0$*, every* $a_n \to a_\nabla \in \mathbb{A}_\nabla$ *and every* $a_n' \to a_0 \in \mathbb{A}_0$ *with* $a_n + t_n a_n' \in \mathbb{A}_f \; \forall n \in \mathbb{N}$.

The relation between the concepts of *continuous* and *uniform* Hadamard differentiability is presented below.

Lemma C.16. (Uniform and Continuous Differentiability) [van der Vaart (1995, Lemma 3.9.7, p.375)] *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *be topological vector spaces and* $\mathbb{A}_f \subset \mathbb{A}$ *be convex. Let* $f : \mathbb{A}_f \to \mathbb{B}$ *be continuously Hadamard differentiable in a neighborhood of* a_∇*, tangentially to* $\mathbb{A}_0$ *(Definition C.14). Then* f *is uniformly differentiable along every sequence* $a_n \to a_\nabla$ *tangentially to* $\mathbb{A}_0$ *(Definition C.15).*

C.3 Novel Differentiability Concepts

The concepts of differentiability introduced in this section are simple modifications of the uniform Hadamard differentiability introduced in Section C.2. As we shall see, our main objective is to deliver smoothness results that hold uniformly over sequences of functions.

First, let us start by introducing two slight modifications of the uniform differentiability of Definition C.15 (from now on called *uniform differentiability of the first kind*). We shall refer to these variants as *uniform differentiability of the second kind* and *uniform differentiability of the third kind*. The difference lies only in working with either $\nabla_{\mathbb{A}_0} f(a_\nabla, a_0)$ or $\nabla_{\mathbb{A}_0} f(a_n, a_0)$ or even $\nabla_{\mathbb{A}_0} f(a_n, a_n')$.

Definition C.17. (Uniform Differentiability of the Second Kind) *In the context of Definition C.15, the function* f *is said to be uniformly Hadamard differentiable of the second kind, along every sequence* $a_n \to a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$*, tangentially to* $\mathbb{A}_0$*, iff*

$$t_n^{-1}\Big(f(a_n + t_n a_n') - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_n, a_0)\Big) \to 0$$

[5] Convexity is required here only to ensure the appropriate convergence of derivatives.

holds for every sequence $t_n \to 0$*, every* $a_n \to a_\nabla \in \mathbb{A}_\nabla$ *and every* $a'_n \to a_0 \in \mathbb{A}_0$ *with* $a_n + t_n a'_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$*.*

Definition C.18. (Uniform Differentiability of the Third Kind) *In the context of Definition C.15, the function* f *is said to be uniformly Hadamard differentiable of the third kind, along every sequence* $a_n \to a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$*, tangentially to* $\mathbb{A}_0$*, iff*

$$t_n^{-1}\Big(f(a_n + t_n a'_n) - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_n, a'_n)\Big) \to 0$$

holds for every sequence $t_n \to 0$*, every* $a_n \to a_\nabla \in \mathbb{A}_\nabla$ *and every* $a'_n \to a_0 \in \mathbb{A}_0$ *with* $a_n + t_n a'_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$*.*

Remark C.19. *When* $\mathbb{A}$ *and* $\mathbb{B}$ *are equipped with norms, then it is trivial to show that uniform Hadamard differentiability of the third kind satisfies the equivalent representation,*

$$\|f(a_n + a'_n) - f(a_n) - \nabla f(a_n, a'_n)\|_{\mathbb{B}} = o(\|a'_n\|_{\mathbb{A}}) \quad as \quad \|a'_n\| \to 0,$$

for every $a_n \to a_\nabla$*.*

Remark C.20. *Inspection of the conditions and proof of Lemma C.16 in van der Vaart and Wellner (1996, Lemma 3.9.7) reveals immediately that the continuous Hadamard differentiability in Definition C.14 implies also the uniform Hadamard differentiability of the second and third kinds. As a result, the representation of Remark C.19 above is also available; see also Reeds (1976) and Gill (1986).*

Let us now introduce two concepts of smoothness that hold over sequences of functions. Due to its greater simplicity, we start with the concept of *Hadamard sequence* which allows us to derive the convergence in distribution of the SNPII estimator.

Definition C.21. (Hadamard Sequence) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces. Let* $\{t_n\} \subset \mathbb{R}$ *be a vanishing sequence* $t_n \to 0$ *and* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *be Hadamard differentiable at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$ *with derivative* $\nabla_{\mathbb{A}_0} f(a_\nabla) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$*. A sequence of functions* $\{f_n\}_{n\in\mathbb{N}}$ *with* $f_n : \mathbb{A}_f \to \mathbb{B} \;\; \forall n \in \mathbb{N}$*, is said to be a* $(1/t_n)$*-Hadamard sequence w.r.t.* f *at* $a_\nabla \in \mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$ *if,*

$$\frac{f_n(a_\nabla + t_n a_n) - f(a_\nabla)}{t_n} \to \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)$$

holds for every $a_n \to a_0 \in \mathbb{A}_0$*.*

Remark C.22. *In the definition of* Hadamard sequence*, the bounded linear operator* $\nabla_{\mathbb{A}_0} f(a_\nabla) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *is restricted to be the Hadamard derivative of* f*. The importance of this restriction is made obvious by statistical applications such as the Delta method that can be found below.*

The second concept of interest is that of *Hadamard equi-differentiability* of a sequence of functions. This concept is useful in deriving conditions for the convergence rate of the SNPII estimator.

Definition C.23. (Hadamard Equi-Differentiability) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces. A sequence* $\{f_n\}_{n\in\mathbb{N}}$ *of Hadamard differentiable functions* $f_n : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ $\forall n \in \mathbb{N}$ *with derivative at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *denoted* $\nabla_{\mathbb{A}_0} f_n(a_\nabla, \cdot) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *is said to be Hadamard equi-differentiable at* a_∇ *tangentially to* $\mathbb{A}_0$ *if,*

$$\frac{f_n(a_\nabla + t_n a_n) - f_n(a_\nabla) - \nabla f_n(a_\nabla, a_0)}{t_n} \to 0$$

for every $t_n \to 0$ *and every* $a_n \to a_0$ *with* $a_\nabla + t_n a_n \in \mathbb{A}_f$ $\forall n \in \mathbb{N}$.

It is important to note that the concept of *Hadamard equi-differentiability* admits uniform counterparts of the *first*, *second* and *third kinds*. The definitions follow.

Definition C.24. (Uniform Hadamard Equi-Differentiability) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces. A sequence* $\{f_n\}_{n\in\mathbb{N}}$ *of Hadamard differentiable functions* $f_n : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ $\forall n \in \mathbb{N}$ *with derivative at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *denoted* $\nabla_{\mathbb{A}_0} f_n(a_\nabla, \cdot) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *is said to be uniformly Hadamard equi-differentiable of the first kind, along every sequence* $a_n \to a_\nabla$ *tangentially to* $\mathbb{A}_0$*, if*

$$\frac{f_n(a_n + t_n a'_n) - f_n(a_n) - \nabla f_n(a_\nabla, a_0)}{t_n} \to 0$$

for every $t_n \to 0$*, every* $a_n \to a_\nabla$*, and every* $a'_n \to a_0$ *with* $a_n + t_n a'_n \in \mathbb{A}_f$ $\forall n \in \mathbb{N}$*. The sequence* $\{f_n\}_{n\in\mathbb{N}}$ *is said to be uniformly Hadamard equi-differentiable of the second kind, along every sequence* $a_n \to a_\nabla$ *tangentially to* $\mathbb{A}_0$*, if for every* $n \in \mathbb{N}$*,* f_n *is Hadamard differentiable at* a_n $\forall n \in \mathbb{N}$*, tangentially to* $\mathbb{A}_0$*, with derivative denoted* $\nabla f_n(a_n, \cdot) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ *and*

$$\frac{f_n(a_n + t_n a'_n) - f_n(a_n) - \nabla f_n(a_n, a_0)}{t_n} \to 0$$

for every $t_n \to 0$*, every* $a_n \to a_\nabla$*, and every* $a'_n \to a_0$ *with* $a_n + t_n a'_n \in \mathbb{A}_f$ $\forall n \in \mathbb{N}$*. Finally, the sequence* $\{f_n\}_{n\in\mathbb{N}}$ *is said to be uniformly Hadamard equi-differentiable of the third kind, along every sequence at* $a_n \to a_\nabla \in \mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0$*, if*

$$\frac{f_n(a_n + t_n a'_n) - f_n(a_n) - \nabla f_n(a_n, a_n)}{t_n} \to 0$$

for every $t_n \to 0$*, every* $a_n \to a_\nabla$*, and every* $a'_n \to a_0$ *with* $a_n + t_n a'_n \in \mathbb{A}_f$ $\forall n \in \mathbb{N}$.

Remark C.25. *When* $\mathbb{A}$ *and* $\mathbb{B}$ *are equipped with norms, then it is trivial to show that uniform Hadamard differentiability of the third kind satisfies the equivalent representation,*

$$\|f_n(a_n + a'_n) - f_n(a_n) - \nabla f_n(a_n, a'_n)\|_{\mathbb{B}} = o(\|a'_n\|_{\mathbb{A}}) \quad as \quad \|a'_n\| \to 0,$$

for every $a_n \to a_\nabla$.

Below we proceed to provide a couple of results on *uniform differentiability* and to characterize *Hadamard sequences* and *uniformly Hadamard equi-differentiable sequences*. In particular, Section C.3.1 analyses the *uniform differentiability* of compositions and product operators. Section C.3.2 shows how *Hadamard sequences* relate to other smoothness concepts, provides alternative sets of sufficient conditions for a sequence of operators to be Hadamard, and finally, derives an adapted delta method that holds for *Hadamard sequences*. Section C.3.3 provides sufficient conditions for a sequence of operators to be *uniformly Hadamard equi-differentiable* and establishes a representation that is useful for the convergence theorem of the SNPII estimator.

C.3.1 Some Results on Uniform Differentiability

As we shall now see, the *uniform Hadamard differentiability of compositions* follows from similar properties on its components.

Proposition C.26. (Uniform Differentiability of Compositions) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$, $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *and* $(\mathbb{C}, \mathcal{T}_{\mathbb{C}})$ *be topological vector spaces. Let* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *be uniformly Hadamard differentiable of the first (second) [third] kind along every sequence* $a_n \to a_\nabla$. *Let* $g : \mathbb{B}_g \subseteq \mathbb{B} \to \mathbb{C}$ *be uniformly Hadamard differentiable of the first (second) [third] kind along the sequence* $b_n := f(a_n) \to b_\nabla := f(a_\nabla)$. *Then the composition* $h := g \circ f \; \forall n \in \mathbb{N}$ *is uniformly Hadamard differentiable of the first (second) [third] kind along every sequence* $a_n \to a_\nabla$.

Proof. By definition, h is uniformly Hadamard differentiable of the first kind along every sequence $a_n \to a_\nabla$ if

$$h(a_n + t_n a'_n) - h(a_n) - t_n \nabla h(a_\nabla, a_0) = o(t_n)$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Now, note that by the chain rule (Lemma C.9)

$$\begin{aligned} &h(a_n + t_n a'_n) - h(a_n) - t_n \nabla h(a_\nabla, a_0) \\ &\quad = g \circ f(a_n + t_n a'_n) - g \circ f(a_n) - t_n \nabla g \circ f(a_\nabla, a_0) \\ &\quad = g \circ f(a_n + t_n a'_n) - g \circ f(a_n) - t_n \nabla g(f(a_\nabla), \nabla f(a_\nabla, a_0)). \end{aligned}$$

Define $b_n := f(a_n)$ and $b_\nabla := f(a_\nabla)$ and also $b'_n := t_n^{-1}\Big(f(a_n + t_n a'_n) - f(a_n)\Big)$ and $b_0 := \nabla f(a_\nabla, a_0)$. Then,

$$\begin{aligned} &h(a_n + t_n a'_n) - h(a_n) - t_n \nabla h(a_\nabla, a_0) \\ &\qquad = g(b_n + t_n b'_n) - g(b_n) - t_n \nabla g(b_n, b_0) \end{aligned}$$

Now note that uniform Hadamard differentiability of the third kind of g along sequences $b_n \to b_\nabla$ implies,

$$g_n(b_n + t_n b'_n) - g_n(b_n) - t_n \nabla g_n(b_\nabla, b_0) = o(t_n)$$

for every $t_n \to 0$, every $b_n \to b_\nabla$ and every $b'_n \to b_0 \in \mathbb{B}$ with $b_n + t_n b'_n \in \mathbb{B}_g \; \forall n \in \mathbb{N}$. Uniform Hadamard differentiability of the first kind of f along sequences $a_n \to a_\nabla$ implies,

$$b'_n \to b_0 \;\; \Leftrightarrow \;\; t_n^{-1}\Big(f(a_n + t_n a'_n) - f(a_n)\Big) \to \nabla f(a_\nabla, a_0)$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. As a result,

$$h(a_\nabla + t_n a_n) - h(a_\nabla) - \nabla h(a_\nabla, a_0) = g(b_n + t_n b'_n) - g(b_n) - \nabla g(b_n, b'_n) = o(t_n)$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Similar reasoning applies to uniform differentiability of the second and third kinds. □

In the context of the product topology, it is important to note that an extension of Proposition C.11 to the various forms of uniform differentiability introduced above is readily available.

Proposition C.27. (Uniform Differentiability with Product Topology) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ *be topological vector spaces for every* i *in some countable index set* $\mathbb{I}$ *and* $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *be the product space* $\mathbb{B} = \times_{i \in \mathbb{I}} \mathbb{B}_i$ *with product topology* $\mathcal{T}_\mathbb{B}$*. Then, a map* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *is uniformly Hadamard differentiable of the first, second or third kinds, along every sequence* $a_n \to a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$ *if and only if the coordinate projection* $\pi_i f : \mathbb{A}_f \to \mathbb{B}_i$ *is also uniformly Hadamard differentiable of the first, second or third kinds respectively, along every sequence* $a_n \to a_\nabla \in \mathbb{A}_\nabla$ *tangentially to* $\mathbb{A}_0 \in \mathbb{A}$*, for every* $i \in \mathbb{I}$*.*

Proof. The desired result follows essentially by the same argument as in the proof of Proposition C.11. By definition, f is uniformly Hadamard differentiable of the first kind, along every sequence $a_n \to a_\nabla \in \mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0 \subseteq \mathbb{A}$ if and only if there exists a continuous linear functional $\nabla_{\mathbb{A}_0} f(a_\nabla) : \mathbb{A}_0 \to \mathbb{B}$ such that, every sequence $\Big\{b_n(t_n, a_n, a'_n)\Big\}_{T \in \mathbb{N}} \subset \mathbb{B}$ defined as,

$$b_n(t_n, a_n, a'_n) := \frac{f(a_n + t_n a'_n) - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)}{t_n}$$

converges to zero, for every $t_n \to 0$, every sequence $a_n \to a_\nabla$ and every sequence $a_n \to a_0 \in \mathbb{A}_0$ with $a_n + t_n a'_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Now, by Corollary A.16 given the product topology $\mathcal{T}_\mathbb{B}$ on $\mathbb{B}$, convergence of the sequence $b_n(t_n, a_n, a'_n) \to 0$ on the product space $\mathbb{B}$ occurs if and only if its coordinate projections $\pi_i b_n(t_n, a_n, a'_n)$ also converge $\pi_i b_n(t_n, a_n, a'_n) \to 0$ in $\mathbb{B}_i$ for every $i \in \mathbb{I}$. By linearity and continuity of the coordinate projection (Lemma A.15 and Proposition A.50) and the definition of Hadamard derivative, it follows immediately that $\nabla \pi_i(\boldsymbol{\beta}) = \pi_i \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ for every $i \in \mathbb{I}$. By the chain rule we thus obtain,

$$\nabla_{\mathbb{A}_0} \pi_i f(a_\nabla, a_0) = \pi_i \nabla_{\mathbb{A}_0} f(a_\nabla, a_0).$$

As a result, we then have that $b_n(t_n, a_n, a'_n) \to 0$ if and only if,

$$\begin{aligned}\pi_i b_n(t_n, a_n, a'_n) :&= \pi_i \left(\frac{f(a_n + t_n a'_n) - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)}{t_n} \right) \\ &= \frac{\pi_i f(a_n + t_n a'_n) - \pi_i f(a_n) - t_n \nabla_{\mathbb{A}_0} \pi_i f(a_\nabla, a_0)}{t_n} \to 0 \text{ for every } i \in \mathbb{I}. \end{aligned} \tag{C.5}$$

Finally, since by Lemma B.5 a composition of linear maps is linear, and by lemma A.29 a composition of continuous maps is continuous, $\pi_i \circ \nabla_{\mathbb{A}_0} f(a_\nabla) \equiv \nabla_{\mathbb{A}_0} \pi_i f$ is a continuous linear map on $\mathbb{A}_0$. This implies, by definition, that the convergence in (C.5) above holds if and only if $\pi_i f$ is uniformly Hadamard differentiable of the first kind along every sequence $a_n \to a_\nabla$ for every $i \in \mathbb{I}$. The same argument applies to sequences, $b_n(t_n, a_n, a'_n) := t_n^{-1}(f(a_n + t_n a'_n) - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_n, a_0))$ and $b_n(t_n, a_n, a'_n) := t_n^{-1}(f(a_n + t_n a'_n) - f(a_n) - t_n \nabla_{\mathbb{A}_0} f(a_n, a'_n))$, and hence, also to uniform Hadamard differentiability of the second and third kinds. □

C.3.2 Characterization of Hadamard Sequences

Proposition C.28. (Characterization of t_n^{-1}-Hadamard Sequences) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *be topological vector spaces and* $t_n \to 0$ *be a sequence in* $\mathbb{R}$. *Let* $\{f_n\}_{n\in\mathbb{N}}$ *be a sequence of functions* $f_n : \mathbb{A}_f \subset \mathbb{A} \to \mathbb{B}\ \forall n \in \mathbb{N}$ *satisfying*

$$f_n(a_\nabla + t_n a_n) - f(a_\nabla + t_n a_n) = o(t_n) \quad \textit{for every } a_n \to a_0$$

such that $a_\nabla + t_n a_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$, *for some limit function* $f : \mathbb{A}_f \to \mathbb{B}$ *that is Hadamard differentiable at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$. *Then* $\{f_n\}_{n\in\mathbb{N}}$ *is a* t_n^{-1}*-Hadamard sequence w.r.t.* f *at* a_∇ *tangentially to* $\mathbb{A}_0$.

Proof. The desired result follows by having,

$$\begin{aligned} f_n(a_\nabla + t_n a_n) - f(a_\nabla) - \nabla f(a_\nabla, t_n a_n) &= \Big[f_n(a_\nabla + t_n a_n) - f(a_\nabla + t_n a_n) \Big] \\ &\quad + \Big[f(a_\nabla + t_n a_n) - f(a_\nabla) - \nabla f(a_\nabla, t_n a_n) \Big] \\ &= o(t_n) + o(t_n) = o(t_n), \end{aligned}$$

for every $a_n \to a_0 \in \mathbb{A}_0$ and $t_n \to 0$ such that $a_\nabla + t_n a_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$.

□

Remark C.29. *When* $\mathbb{B}$ *is equipped with a metric, then sufficient conditions for (C.28) above include* $\sup_{a \in S(\boldsymbol{\theta}_0, \delta_n)} \delta_\mathbb{B}(f_n(a), f(a)) = o(t_n)$ *for every sequence* $\delta_n = O(t_n)$, *as well as the simpler albeit considerably more restrictive uniform convergence,* $\sup_{a \in \mathbb{A}^*} \delta_\mathbb{B}(f_n(a), f(a)) = o(t_n)$ *for every compact subset* $\mathbb{A}^* \subseteq \mathbb{A}$.

Proposition C.30. (Characterization of t_n^{-1}-Hadamard Sequences) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces and let* $f_n : \mathbb{A}_f \subset \mathbb{A} \to \mathbb{B}$ *be Hadamard equi-differentiable at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f \; \forall n \in \mathbb{N}$ *with tangential derivative* $\nabla_{\mathbb{A}_0} f_n(a_\nabla) : \mathbb{A}_0 \to \mathbb{B}$ *satisfying* $\nabla_{\mathbb{A}_0} f_n(a_\nabla, a_0) \to \nabla_{\mathbb{A}_0} f(a_\nabla, a_0)$ *as* $n \to \infty \, \forall a_0 \in \mathbb{A}_0$ *where* $\nabla_{\mathbb{A}_0} f(a_\nabla)$ *denotes the tangential derivative of some limit Hadamard differentiable function* $f : \mathbb{A}_f \to \mathbb{B}$. *Furthermore, suppose that* $f_n(a_\nabla) - f(a_\nabla) = o(t_n)$. *Then,* $\{f_n\}_{n\in\mathbb{N}}$ *is a* t_n^{-1}*-Hadamard sequence w.r.t.* f *at* a_∇ *tangentially to* $\mathbb{A}_0$.

Proof. For every $a_n \to a_0 \in \mathbb{A}_0$ and $t_n \to 0$ such that $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$,

$$\begin{aligned}\frac{f_n(a_\nabla + t_n a_n) - f(a_\nabla)}{t_n} - \nabla_{\mathbb{A}_0} f(a_\nabla, a_0) =& \frac{f_n(a_\nabla + t_n a_n) - f_n(a_\nabla)}{t_n} - \nabla_{\mathbb{A}_0} f_n(a_\nabla, a_0) \\ &+ \frac{f_n(a_\nabla) - f(a_\nabla)}{t_n} + \nabla_{\mathbb{A}_0} f_n(a_\nabla, a_0) - \nabla_{\mathbb{A}_0} f(a_\nabla, a_0).\end{aligned}$$

Finally, note that $(1/t_n)(f_n(a_\nabla + t_n a_n) - f_n(a_\nabla)) - \nabla_{\mathbb{A}_0} f_n(a_\nabla, a_0) = o(1)$ follows from the Hadamard equi-differentiability of $f_n \; \forall n \in \mathbb{N}$ and the convergence to a differentiable limit and $(1/t_n)(f_n(a_\nabla) - f(a_\nabla)) = o(1)$ and $\nabla_{\mathbb{A}_0} f_n(a_\nabla, a_0) - \nabla_{\mathbb{A}_0} f(a_\nabla, a_0) = o(1)$ by the conditions imposed. □

The following proposition is especially useful in the SNPII setting where the properties of the criterion function Q_T and its limit Q_∞ are derived from the properties of the composition of μ_T and μ_∞ with $\Delta_{T,S}$ and Δ_∞ respectively. The assumptions on the spaces are thus especially tailored to the SNPII estimator. In particular, the proposition below postulates in its assumptions a 'domain' space consisting of normed vector space (resembling Θ), an 'intermediate' space taking the form of a vector space (resembling $\mathcal{B}$) equipped with a difference metric like $\delta_{\mathcal{B}}$ (Definition A.40 and Remark A.41), and an 'image' space that is also a normed vector space (resembling $\mathbb{R}$).

Proposition C.31. (Characterization of t_n^{-1}-Hadamard Sequences for Composite Operators) *Let* $(\mathbb{A}, \|\cdot\|_{\mathbb{A}})$ *and* $(\mathbb{C}, \|\cdot\|_{\mathbb{C}})$ *be normed vector spaces,* $(\mathbb{B}, \delta_{\mathbb{B}})$ *be a vector space equipped with a difference metric* $\delta_{\mathcal{B}}$*, and* $t_n \to 0$ *be a sequence in* $\mathbb{R}$. *Let* $\{g_n\}_{n\in\mathbb{N}}$ *and* $\{f_n\}_{n\in\mathbb{N}}$ *be sequences of functions* $g_n : \mathbb{B}_g \subseteq \mathbb{B} \to \mathbb{C}$ *and* $f_n : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B} \; \forall n \in \mathbb{N}$ *satisfying,*

$$\|g_n(b_n) - g(b_\nabla)\|_{\mathbb{C}} = O\Big(\nu(\delta_{\mathbb{B}}(b_n, b_\nabla))\Big)$$

for some $\nu : \mathbb{R} \to \mathbb{R}$ *satisfying* $\nu(r_n) \to 0$ *as* $r_n \to 0$, *and every* $b_n \to b_\nabla$ *where* $b_\nabla = f(a_\nabla)$ *and* $g : \mathbb{B}_g \to \mathbb{C}$ *is Hadamard differentiable at* $b_\nabla \in \mathbb{B}_\nabla \subset \mathbb{B}_g$ *tangentially to* $\mathbb{B}_0 \subseteq \mathbb{B}$ *with*

$$\|g(b_n) - g(b_\nabla)\|_{\mathbb{C}} = O\Big(\nu(\delta_{\mathbb{B}}(b_n, b_\nabla))\Big)$$

for every sequence $b_n - b_\nabla \to 0$. *Finally, let*

$$\delta_{\mathbb{B}}\Big(f_n(a_n)\,,\, f(a_\nabla)\Big) = O\Big(\mu(\|a_n - a_\nabla\|_{\mathbb{A}})\Big)$$

for some $\mu : \mathbb{R} \to \mathbb{R}$ *satisfying* $\mu(r_n) \to 0$ *as* $r_n \to 0$, *and every* $a_n \to a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *where* $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ *is Hadamard differentiable at* $a_\nabla \in \mathbb{A}_\nabla$. *If* $\nu(r_n) = o\Big(r_n r'_n/\mu(r'_n)\Big)$ *as* $r_n \to 0$ *with* $r'_n = O(r_n)$, *then* $\{g_n \circ f_n\}_{n\in\mathbb{N}}$ *is a* t_n^{-1}*-Hadamard sequence w.r.t.* $g \circ f$ *at* a_∇ *tangentially to* $\mathbb{A}_0$.

Proof. Recall the first step of the proof of Proposition C.28. For every $a_n \to a_0 \in \mathbb{A}_0$ and $t_n \to 0$ such that $a_\nabla + t_n a_n \in \mathbb{A}_f \,\forall n \in \mathbb{N}$ we have by norm sub-additivity,

$$\begin{aligned}
&\Big\|g_n \circ f_n(a_\nabla + t_n a_n) - g \circ f(a_\nabla) - \nabla(g \circ f)(a_\nabla, a_0)\Big\|_{\mathbb{C}} \\
&\qquad \le \Big\|g_n \circ f_n(a_\nabla + t_n a_n) - g \circ f(a_\nabla)\Big\|_{\mathbb{C}} \\
&\qquad\quad + \Big\|g \circ f(a_\nabla) - g \circ f(a_\nabla + t_n a_n)\Big\|_{\mathbb{C}} \\
&\qquad\quad + \Big\|g \circ f(a_\nabla + t_n a_n) - g \circ f(a_\nabla) - \nabla(g \circ f)(a_\nabla, a_0)\Big\|_{\mathbb{C}}.
\end{aligned}$$

Finally, by the Hadamard differentiability of f and g, the Chain rule (Lemma C.9), the properties of difference metrics, and the convergence conditions above,

$$\begin{aligned}
&\Big\|g_n \circ f_n(a_\nabla + t_n a_n) - g \circ f(a_\nabla) - \nabla(g \circ f)(a_\nabla, a_0)\Big\|_{\mathbb{C}} \\
&\qquad \le O\Big(\nu\Big(\delta_{\mathcal{B}}\Big(f_n(a_\nabla + t_n a_n)\,,\, f(a_\nabla)\Big)\Big)\Big) + o(t_n) \\
&\qquad = O\Big(\nu\Big(O\Big(\mu(\|t_n a_n\|_{\mathbb{A}})\Big)\Big)\Big) + o(t_n) \\
&\qquad = O\Big(\nu\Big(O\Big(\mu(O(t_n))\Big)\Big)\Big) + o(t_n) = o(t_n).
\end{aligned}$$

□

Proposition C.32 provides another result obtained under conditions that are especially tailored for the SNPII estimator. Note in particular the use of a vector space with an asymptotically homogeneous difference metric (Definitions A.40 and A.42, and Remarks A.41 and A.43) that resembles $\mathcal{B}$ with a product metric $\delta_{\mathcal{B}}$.

Proposition C.32. (Characterization of t_n^{-1}-Hadamard Sequences for Operators of Two Arguments) *Let* $(\mathbb{A}, \delta_{\mathbb{A}})$ *and* $(\mathbb{B}, \delta_{\mathbb{B}})$ *be vector spaces equipped with asymptotically homogeneous difference metrics,* $(\mathbb{C}, \|\cdot\|_{\mathbb{C}})$ *be a normed vector space, and* $t_n \to 0$ *be a sequence in* $\mathbb{R}$. *Let* $\{f_n\}_{n\in\mathbb{N}}$ *be a sequence of functions* $f_n : \mathbb{A}_f \times \mathbb{B} \to \mathbb{C} \,\forall n \in \mathbb{N}$ *where* $f_n(a, \cdot) \in \mathbb{L}(\mathbb{B}, \mathbb{C})$ *satisfying,*

$$\sup_{(a,b)\in S(a_\nabla,\epsilon)\times S(b_\nabla,\epsilon')} \Big\|f_n(a, b) - f(a, b)\Big\|_{\mathbb{C}} = o(t_n)$$

for some pair $\epsilon > 0$ *and* $\epsilon' > 0$ *where* $f : \mathbb{A}_f \times \mathbb{B} \to \mathbb{C}$ *with* $f(a, \cdot) : \mathbb{L}(\mathbb{B}, \mathbb{C})$, *satisfies* (i) $f(a, 0) = 0 \,\forall a \in \mathbb{A}_f$; (ii) $f(\cdot, b_\nabla) : \mathbb{B}_g \to \mathbb{C}$ *is Hadamard differentiable*

at $a_\nabla \in \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subset \mathbb{A}$ *and;* (iii) $\sup_{a\in\mathbb{A}_0} \|f(a, b'_n)\|_{\mathbb{B}} = o(\xi_g(\delta_{\mathbb{B}}(b'_n)))$ *for every sequence* $b'_n \to 0$. *If* $\delta_{\mathbb{B}}(b_n, b_\nabla) = O(r_n)$ *with* $r_n = \xi_g^1(t_n)$, *then* $\{f_n(\cdot, b_n)\}_{n\in\mathbb{N}}$ *is a* t_n^{-1}*-Hadamard sequence w.r.t.* $f(\cdot, b_\nabla)$ *at* a_∇ *tangentially to* $\mathbb{A}_0$.

Proof. For every $a_n \to a_0 \in \mathbb{A}_0$ and $t_n \to 0$ such that $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$ we have by norm sub-additivity,

$$\begin{aligned}
&\Big\|f_n(a_\nabla + t_n a_n, b_n) - f(a_\nabla, b_\nabla) - \nabla_{\mathbb{A}_0} f_{a_\nabla}(a_0, b_\nabla)\Big\|_{\mathbb{C}} \\
&\qquad \le \Big\|f_n(a_\nabla + t_n a_n, b_n) - f(a_\nabla + t_n a_n, b_n)\Big\|_{\mathbb{C}} \\
&\qquad\quad + \Big\|f(a_\nabla + t_n a_n, b_n) - f(a_\nabla + t_n a_n, b_\nabla)\Big\|_{\mathbb{C}} \\
&\qquad\quad + \Big\|f(a_\nabla + t_n a_n, b_\nabla) - f(a_\nabla, b_\nabla) - \nabla f_{a_\nabla}(a_0, b_\nabla)\Big\|_{\mathbb{C}}.
\end{aligned}$$

Now, since $t_n a_n \to 0$ it follows that $\exists\, n^* \in \mathbb{N}$ such that $a_\nabla + t_n a_n \in S(a_\nabla, \epsilon) \; \forall n > n^*$ and every $\epsilon > 0$. Furthermore, by the same argument $\exists\, n^* \in \mathbb{N}$ such that $b_n \in S(b_\nabla, \epsilon') \; \forall n > n^*$ and every $\epsilon' > 0$. As a result, for every $n > n^*$,

$$\begin{aligned}
&\Big\|f_n(a_\nabla + t_n a_n, b_n) - f(a_\nabla, b_\nabla) - \nabla f_{a_\nabla}(t_n a_n, b_\nabla)\Big\|_{\mathbb{C}} \\
&\qquad \le \sup_{(a,b)\in S(a_\nabla,\epsilon)\times S(b_\nabla,\epsilon')} \Big\|f_n(a, b) - f(a, b)\Big\|_{\mathbb{C}} \\
&\qquad\quad + \sup_{a\in S(a_\nabla,\epsilon)} \Big\|f(a, b_n - b_\nabla)\Big\|_{\mathbb{C}} + o(t_n) \\
&\qquad = o(t_n) + o\Big(\xi_f(O(r_n))\Big) = o(t_n) + o(t_n) = o(t_n).
\end{aligned}$$

□

Corollary C.33. (Characterization of t_n^{-1}-Hadamard Sequences) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *be a topological vector space,* $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ *be a normed vector space, and* $t_n \to 0$ *be a sequence in* $\mathbb{R}$. *Let* $\{f_n\}_{n\in\mathbb{N}}$ *be a sequence of functions* $f_n : \mathbb{A}_f \subset \mathbb{A} \to \mathbb{B} \; \forall n \in \mathbb{N}$ *satisfying* $\sup_{a\in\mathbb{A}^*} \Big\|f_n(a) - f(a)\Big\|_{\mathbb{B}} = o(t_n)$ *for every compact subset* $\mathbb{A}^* \subseteq \mathbb{A}$ *for some limit function* $f : \mathbb{A}_f \to \mathbb{B}$ *that is Hadamard differentiable at* $\alpha_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subset \mathbb{A}$. *Then* $\{f_n\}_{n\in\mathbb{N}}$ *is a* t_n^{-1}*-Hadamard sequence w.r.t.* f *at* a_∇ *tangentially to* $\mathbb{A}_0$.

The usefulness of *Hadamard sequences* is now revealed with the introduction of an especially designed delta method.

Proposition C.34. (Delta Method for Hadamard Sequences) *Let* $(\mathbb{A}, \mathcal{T}_{\mathbb{A}})$ *and* $(\mathbb{B}, \mathcal{T}_{\mathbb{B}})$ *be topological vector spaces and* $\{f_n\}_{n\in\mathbb{N}}$ *be a* $(1/t_n)$*-Hadamard sequence of measurable maps* $f_n : \mathbb{A}_f \subset \mathbb{A} \to \mathbb{B} \; \forall n \in \mathbb{N}$ *w.r.t. a map* $f : \mathbb{A}_f \subset \mathbb{A} \to \mathbb{B}$ *with measurable Hadamard derivative at* $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ *tangentially to* $\mathbb{A}_0 \subseteq \mathbb{A}$. *Then, if* $\{X_n\}_{n\in\mathbb{N}}$ *is a random sequence satisfying* $(1/t_n)(X_n - a_\nabla) \xrightarrow{d} Z$ *for some tight random element* Z *taking values in* $\mathbb{A}_0$, *it follows that,*

$$(1/t_n)\Big(f_n(X_n) - f(a_\nabla)\Big) \xrightarrow{d} \nabla f(a_\nabla, Z).$$

Proof. Define the sequence of functions $\{g_n\}_{n\in\mathbb{N}}$ according to,

$$g_n(a_n) = \frac{f_n(a_\nabla + t_n a_n) - f(a_\nabla)}{t_n} \; \forall n \in \mathbb{N}.$$

$g_n(a_n) \to \nabla f(a_\nabla, a_0)$ for every $a_n \to a_0 \in \mathbb{A}_0$ follows immediately from $\{f_n\}_{n\in\mathbb{N}}$ being a $(1/t_n)$-Hadamard sequence at a_∇. By the ECMT (Lemma A.54) it then follows that $g_n((1/t_n)(X_n - a_\nabla)) \to \nabla f(a_\nabla, Z)$. Finally, note that,

$$\begin{aligned} g_n\Big((1/t_n)(X_n - a_\nabla)\Big) &= \frac{f_n\Big(a_\nabla + t_n(1/t_n)(X_n - a_\nabla)\Big) - f(a_0)}{t_n} \\ &= (1/t_n)\Big(f_n(X_n) - f(a_\nabla)\Big). \end{aligned}$$

□

Proposition C.35. (Delta Method for Continuous Bilinear Sequences) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$, $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *and* $(\mathbb{C}, \mathcal{T}_\mathbb{C})$ *be topological vector spaces and* $\{f_n\}_{n\in\mathbb{N}} \subset \mathbb{L}^2(\mathbb{A}_f \times \mathbb{B}_f, \mathbb{C})$ *be a sequence of bounded bilinear operators* $f_n : \mathbb{A}_f \times \mathbb{B}_f \subset \mathbb{A} \times \mathbb{B} \to \mathbb{C} \; \forall n \in \mathbb{N}$ *satisfying* $f_n(a_n, b_n) \to f(a, b)$ *for every* $(a_n, b_n) \to (a, b) \in \mathbb{A}_f \times \mathbb{B}_f$ *where* $f \in \mathbb{L}^2(\mathbb{A}_f \times \mathbb{B}_f, \mathbb{C})$. *Furthermore, let* $\{X_n\}_{n\in\mathbb{N}}$ *is a random sequence satisfying* $(1/t_n)(X_n - a_0) \xrightarrow{d} Z$ *for some random element* Z *taking values in* $\mathbb{A}_0 \subseteq \mathbb{A}$ *and* $\{b_n\}_{n\in\mathbb{N}}$ *be a sequence satisfying* $b_n \to b_0 \in \mathbb{B}_f \subseteq \mathbb{B}$. *Then,*

$$(1/t_n)\Big(f_n(X_n, b_n) - f(a_0, b_0)\Big) \xrightarrow{d} \nabla_{\mathbb{A}_0 \times \mathbb{B}_f} f(Z, b_0), \quad as \;\; T \to \infty,$$

if either $f_n(a_0, b_n) - f(a_0, b_0) = o(t_n)$ *or alternatively* $a_0 = 0$ *and* $f(a_0, b_0) = 0$.[6]

Proof. In the latter case of $a_0 = 0$ and $f(a_0, b_0) = 0$ it follows immediately that

$$(1/t_n)\Big(f_n(X_n, b_n) - f(a_0, b_0)\Big) = (1/t_n) f_n(X_n, b_n) = f_n((1/t_n)X_n, b_n) \xrightarrow{d} f(Z, b_0),$$

where the last step follows from the fact that $f_n(a_n, b_n) \to f(a, b) \; \forall (a_n, b_n) \to (a, b)$ and an application of the ECMT (Proposition A.54). Finally, in the former case the desired result is obtained immediately by noting that by Proposition C.30, the sequence of bilinear maps satisfying $f_n(a_0, b_n) - f(a_0, b_0) = o(t_n)$ is immediately a Hadamard sequence. □

[6]Note here that while f is defined on the restricted domain $\mathbb{A}_f \times \mathbb{B}_f$, its derivative $\nabla_{\mathbb{A}_0 \times \mathbb{B}_0} f$ is defined on $\mathbb{A}_0 \times \mathbb{B}_f$. In essence, $\nabla_{\mathbb{A}_0 \times \mathbb{B}_0} f$ can be seen as a bilinear extension of the bilinear map f from $\mathbb{A}_f \times \mathbb{B}_f$ to $\mathbb{A}_0 \times \mathbb{B}_f$. This is important because while the bilinear map might be defined on a compact set $\mathbb{A}_f \times \mathbb{B}_f$, we must allow $\mathbb{A}_0$ to be unbounded so that Z might have an unbounded support (e.g. to be Gaussian). So just like in the usual case we allow the derivative to be defined on a larger set. Compactness of $\mathbb{A}_f \times \mathbb{B}_f$ is often important to derive the uniform convergence of the sequence of maps $\{f_n\}$ and thus obtain $f_n(a_n, b_n) \to f(a, b)$ for every $(a_n, b_n) \to (a, b)$.

Remark C.36. *Note that to obtain (3.8) the case of interest to us in Proposition (C.35) is the one that assumes $a_0 = 0$ and $f(a_0, b_0) = 0$, since indeed, in (3.8) we have,*

$$\sqrt{T}\Big[\nabla\mu_T(\Delta_{T,S}(\boldsymbol{\theta}_0), \nabla\Delta_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T})) - \nabla\mu_\infty(\Delta_\infty(\boldsymbol{\theta}_0), \nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta))\Big]$$

with $\Delta_\infty(\boldsymbol{\theta}_0) = 0$ and $\mu_\infty(\Delta_\infty(\boldsymbol{\theta}_0), \nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta)) = 0$ by construction. Hence, (3.8) is obtained if $\nabla\Delta_{T,S}(\boldsymbol{\theta}_0, \mathbb{S}_{\Theta_T}) - \nabla\Delta_\infty(\boldsymbol{\theta}_0, \mathbb{S}_\Theta) = o_p(1)$ and $\sup_{(\boldsymbol{\beta},\boldsymbol{\beta}')\in\mathcal{B}\times\mathcal{B}} \Big|\nabla\mu_T(\boldsymbol{\beta}, \boldsymbol{\beta}') - \nabla\mu_\infty(\boldsymbol{\beta}, \boldsymbol{\beta}')\Big| = o(1)$.

Proposition C.3.1. (Bilinear Hadamard Sequences) *Let $(\mathbb{A}, \mathcal{T}_\mathbb{A})$, $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ and $(\mathbb{C}, \mathcal{T}_\mathbb{C})$ be topological vector spaces and $\{f_n\}_{n\in\mathbb{N}} \subset \mathbb{L}^2(\mathbb{A}_f \times \mathbb{B}_f, \mathbb{C})$ be a sequence of bounded bilinear operators $f_n : \mathbb{A}_f \times \mathbb{B}_f \subset \mathbb{A} \times \mathbb{B} \to \mathbb{C}\ \forall n \in \mathbb{N}$ satisfying $f_n(a_n, b_n) \to f(a, b)$ for every $(a_n, b_n) \to (a, b) \in \mathbb{A}_f \times \mathbb{B}_f$ where $f \in \mathbb{L}^2(\mathbb{A}_f \times \mathbb{B}_f, \mathbb{C})$. If $f(a_\nabla) = 0$ then $\{f_n\}_{n\in\mathbb{N}}$ is a $(1/t_n)$-Hadamard sequence at a_∇ for every $t_n \to 0$.*

Proof. For every $a_n \to a \in \mathbb{A}$ and $t_n \to 0$ such that $a_\nabla + t_n a_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$,

$$\frac{f_n(a_\nabla + t_n a_n) - f(a_\nabla)}{t_n} - f(a_\nabla, a) = f_n(a_\nabla/t_n + a_n) - f(a) \to 0,$$

where the equality follows immediately from $f(a_\nabla) = 0$ the bilinearity of f_n and the final convergence follows from having $f_n(a_n, b_n) \to f(a, b)$ for every $(a_n, b_n) \to (a, b) \in \mathbb{A}_f \times \mathbb{B}_f$. □

Let us finally observe that the concept of *Hadamard sequence* is amenable to the product topology 'treatment' that we have explored in Propositions C.11 and C.11.

Proposition C.37. (Hadamard Sequence with Product Topology) *Let $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ and $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ be topological vector spaces for every i in some countable index set $\mathbb{I}$ and $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ be the product space $\mathbb{B} = \times_{i\in\mathbb{I}}\mathbb{B}_i$ with product topology $\mathcal{T}_\mathbb{B}$. Let $\{t_n\} \subset \mathbb{R}$ be a vanishing sequence $t_n \to 0$ and $f : \mathbb{A}_f \subseteq \mathbb{A} \to \mathbb{B}$ be Hadamard differentiable at $a_\nabla \in \mathbb{A}_\nabla \subset \mathbb{A}_f$ tangentially to $\mathbb{A}_0 \subseteq \mathbb{A}$ with derivative $\nabla_{\mathbb{A}_0} f(a_\nabla) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$. Then, a sequence of functions $\{f_n\}_{n\in\mathbb{N}}$ with $f_n : \mathbb{A}_f \to \mathbb{B}\ \ \forall n \in \mathbb{N}$, is a $(1/t_n)$-Hadamard sequence w.r.t. f at $a_\nabla \in \mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0$ if and only if the coordinate projections $\pi_i f_n : \mathbb{A}_f \to \mathbb{B}_i$ are $(1/t_n)$-Hadamard sequences w.r.t. $\pi_i f$ at $a_\nabla \in \mathbb{A}_\nabla$ tangentially to $\mathbb{A}_0$, for every $i \in \mathbb{I}$.*

Proof. By definition, $\{f_n\}$ is a $(1/t_n)$-Hadamard sequence w.r.t. f at a_∇, tangentially to $\mathbb{A}_0$, if and only if every sequence $\Big\{b_n(t_n, a_n)\Big\}_{T\in\mathbb{N}} \subset \mathbb{B}$ defined as,

$$b_n(t_n, a_n) := \frac{f_n(a_\nabla + t_n a_n) - f(a_\nabla) - t_n \nabla f(a_\nabla, a_0)}{t_n}$$

converges to zero for every $a_n \to a_0 \in \mathbb{A}_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$. Now, by Corollary A.16 given the product topology $\mathcal{T}_\mathbb{B}$ on $\mathbb{B}$, convergence of the sequence

$b_n(t_n, a_n) \to 0$ on the product space $\mathbb{B}$ occurs if and only if its coordinate projections $\pi_i b_n(t_n, a_n)$ vanish $\pi_i b_n(t_n, a_n) \to 0$ in $\mathbb{B}_i$ for every $i \in \mathbb{I}$. By linearity and continuity of the coordinate projection (Lemma A.15 and Proposition A.50) and the definition of Hadamard derivative, it follows immediately that $\nabla \pi_i(\boldsymbol{\beta}) = \pi_i \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ for every $i \in \mathbb{I}$. By the chain rule we thus obtain,

$$\nabla_{\mathbb{A}_0} \pi_i f(a_\nabla, a_0) = \pi_i \nabla_{\mathbb{A}_0} f(a_\nabla, a_0).$$

As a result, we then have that $b_n(t_n, a_n) \to 0$ if and only if,

$$\begin{aligned} \pi_i b_n(t_n, a_n) :&= \pi_i \left(\frac{f_n(a_\nabla + t_n a_n) - f(a_\nabla) - t_n \nabla f(a_\nabla, a_0)}{t_n} \right) \\ &= \frac{\pi_i f_n(a_\nabla + t_n a_n) - \pi_i f(a_\nabla) - t_n \nabla \pi_i f(a_\nabla, a_0)}{t_n} \to 0 \text{ for every } i \in \mathbb{I}. \end{aligned} \tag{C.6}$$

Finally, since by Lemma B.5 a composition of linear maps is linear, and by lemma A.29 a composition of continuous maps is continuous, $\pi_i \circ \nabla_{\mathbb{A}_0} f(a_\nabla)$ is a continuous linear map on $\mathbb{A}_0$. This implies, by definition, that the convergence in (C.6) above holds if and only if $\pi_i f$ is uniformly Hadamard differentiable of the first kind along every sequence $a_n \to a_\nabla$ for every $i \in \mathbb{I}$. □

C.3.3 Results on Uniform Hadamard Equi-differentiability

Proposition C.38. (Equi-Differentiability of Compositions) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{C}, \mathcal{T}_\mathbb{C})$ *be topological vector spaces, and* $(\mathbb{B}, \delta_\mathbb{B})$ *be a vector space equipped with a difference metric* $\delta_\mathbb{B}$. *Let the sequence* $\{f_n\}_{n\in\mathbb{N}}$ *of maps* $f_n : \mathbb{A}_f \to \mathbb{B}_g$ *be Hadamard equi-differentiable at* a_∇ *and satisfy* $f_n(a_\nabla) \to f(a_\nabla)$. *Let the sequence* $\{g_n\}_{n\in\mathbb{N}}$ *of maps* $g_n : \mathbb{B}_g \to \mathbb{C}$ *be uniformly Hadamard equi-differentiable of the third kind along the sequence* $b_n := f_n(a_\nabla) \to b_\nabla := f(a_\nabla)$. *Finally, let* $\nabla g_n(b_n, \cdot) \in \mathbb{L}(\mathbb{B}, \mathbb{C})$ *satisfy* $\nabla g_n(b_n, b_n') = O(\delta_\mathbb{B}(b_n'))$ *for every* $b_n' \to 0_\mathbb{B}$. *Then the composition sequence* $\{h_n\}_{n\in\mathbb{N}}$ *with* $h_n := g_n \circ f_n \; \forall n \in \mathbb{N}$ *is Hadamard equi-differentiable at* a_∇.

Proof. By definition, $\{h_n\}_{n\in\mathbb{N}}$ is Hadamard equi-differentiable at a_∇ if

$$h_n(a_\nabla + t_n a_n) - h_n(a_\nabla) - t_n \nabla h_n(a_\nabla, a_0) = o(t_n)$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Now, note that by the chain rule, (Lemma C.9)

$$\begin{aligned} &h_n(a_\nabla + t_n a_n) - h_n(a_\nabla) - t_n \nabla h_n(a_\nabla, a_0) \\ &\quad = g_n \circ f_n(a_\nabla + t_n a_n) - g_n \circ f_n(a_\nabla) - t_n \nabla g_n \circ f_n(a_\nabla, a_0) \\ &\quad = g_n \circ f_n(a_\nabla + t_n a_n) - g_n \circ f_n(a_\nabla) - t_n \nabla g_n(f_n(a_\nabla), \nabla f_n(a_\nabla, a_0)). \end{aligned}$$

Define $b_n := f_n(a_\nabla)$, $b'_n := t_n^{-1}\Big(f_n(a_\nabla + t_n a_n) - f_n(a_\nabla)\Big)$. Then,

$$\begin{aligned} h_n(a_\nabla + t_n a_n) - h_n(a_\nabla) - t_n \nabla h_n(a_\nabla, a_0) \\ = g_n(b_n + t_n b'_n) - g_n(b_n) - t_n \nabla g_n(b_n, b'_n) \\ + \nabla g_n(b_n, t_n(b'_n - \nabla f_n(a_\nabla, a_0))). \end{aligned}$$

Now note that uniform Hadamard equi-differentiability of the third kind of $\{g_n\}$ implies,

$$g_n(b_n + t_n b'_n) - g_n(b_n) - t_n \nabla g_n(b_n, b'_n) = o(t_n)$$

for every $t_n \to 0$, every $b_n \to b_\nabla$ and every $b'_n \to b_0 \in \mathbb{B}$ with $b_n + t_n b'_n \in \mathbb{B}\ \forall n \in \mathbb{N}$. Hadamard equi-differentiability of $\{f_n\}$ implies,

$$t_n(b'_n - \nabla f_n(a_\nabla, a_0)) = f_n(a_\nabla + t_n a_n) - f_n(a_\nabla) - \nabla f_n(a_\nabla, a_0) = o(t_n),$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$. Finally, since

$$\nabla g_n(b_n, t_n(b'_n - \nabla f_n(a_\nabla, a_0))) = O(b'_n - \nabla f_n(a_\nabla, a_0))) = O(o(t_n)) = o(t_n)$$

we have that,

$$\begin{aligned} h_n(a_\nabla + t_n a_n) - h_n(a_\nabla) - \nabla h_n(a_\nabla, a_0) &= g_n(b_n + t_n b'_n) - g_n(b_n) - \nabla g_n(b_n, b'_n) \\ &\quad + \nabla g_n(b_n, b'_n - \nabla f_n(a_\nabla, a_0)) \\ &= o(t_n) + o(t_n) = o(t_n) \end{aligned}$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$. □

Proposition C.39. (Uniform Equi-Differentiability of Compositions) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{C}, \mathcal{T}_\mathbb{C})$ *be topological vector spaces, and* $(\mathbb{B}, \delta_\mathbb{B})$ *be a vector space equipped with a difference metric* $\delta_\mathbb{B}$*. Let the sequence* $\{f_n\}_{n\in\mathbb{N}}$ *of maps* $f_n : \mathbb{A}_f \to \mathbb{B}_g$ *be uniform Hadamard equi-differentiable of the first (second) [third] kind along every sequence* $a_n \to a_\nabla$ *and satisfy* $f_n(a_\nabla) \to f(a_\nabla)$*. Let the sequence* $\{g_n\}_{n\in\mathbb{N}}$ *of maps* $g_n : \mathbb{B}_g \to \mathbb{C}$ *be uniformly Hadamard equi-differentiable of the third kind along the sequence* $b_n := f_n(a_\nabla) \to b_\nabla := f(a_\nabla)$*. Finally, let* $\nabla g_n(b_n, \cdot) \in \mathbb{L}(\mathbb{B}, \mathbb{C})$ *satisfy* $\nabla g_n(b_n, b'_n) = O(\delta_\mathbb{B}(b'_n))$ *for every* $b'_n \to 0_\mathbb{B}$*. Then the composition sequence* $\{h_n\}_{n\in\mathbb{N}}$ *with* $h_n := g_n \circ f_n\ \forall n \in \mathbb{N}$ *is uniformly Hadamard equi-differentiable of the first (second) [third] kind along every sequence* $a_n \to a_\nabla$*.*

Proof. By definition, $\{h_n\}_{n\in\mathbb{N}}$ is uniform Hadamard equi-differentiable of the first kind along every sequence $a_n \to a_\nabla$ if

$$h_n(a_n + t_n a'_n) - h_n(a_n) - t_n \nabla h_n(a_\nabla, a_0) = o(t_n)$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Now, note that by the chain rule (Lemma C.9)

$$\begin{aligned} & h_n(a_n + t_n a'_n) - h_n(a_n) - t_n \nabla h_n(a_\nabla, a_0) \\ & \quad = g_n \circ f_n(a_n + t_n a'_n) - g_n \circ f_n(a_n) - t_n \nabla g_n \circ f_n(a_\nabla, a_0) \\ & \quad = g_n \circ f_n(a_n + t_n a'_n) - g_n \circ f_n(a_n) - t_n \nabla g_n(f_n(a_\nabla), \nabla f_n(a_\nabla, a_0)). \end{aligned}$$

Define $b_n := f_n(a_n)$, $b'_n := t_n^{-1}\Big(f_n(a_n + t_n a'_n) - f_n(a_n)\Big)$. Then,

$$\begin{aligned} & h_n(a_n + t_n a'_n) - h_n(a_n) - t_n \nabla h_n(a_\nabla, a_0) \\ & \qquad = g_n(b_n + t_n b'_n) - g_n(b_n) - t_n \nabla g_n(b_n, b'_n) \\ & \qquad\quad + \nabla g_n(b_n, t_n(b'_n - \nabla f_n(a_\nabla, a_0))). \end{aligned}$$

Now note that uniform Hadamard equi-differentiability of the third kind of $\{g_n\}$ implies,

$$g_n(b_n + t_n b'_n) - g_n(b_n) - t_n \nabla g_n(b_n, b'_n) = o(t_n)$$

for every $t_n \to 0$, every $b_n \to b_\nabla$ and every $b'_n \to b_0 \in \mathbb{B}$ with $b_n + t_n b'_n \in \mathbb{B} \; \forall n \in \mathbb{N}$. uniform Hadamard equi-differentiability of the first kind of $\{f_n\}$ implies,

$$t_n(b'_n - \nabla f_n(a_\nabla, a_0)) = f_n(a_n + t_n a'_n) - f_n(a_n) - \nabla f_n(a_\nabla, a_0) = o(t_n),$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Finally, since

$$\nabla g_n(b_n, t_n(b'_n - \nabla f_n(a_\nabla, a_0))) = O(b'_n - \nabla f_n(a_\nabla, a_0))) = O(o(t_n)) = o(t_n)$$

we have that,

$$\begin{aligned} h_n(a_\nabla + t_n a_n) - h_n(a_\nabla) - \nabla h_n(a_\nabla, a_0) &= g_n(b_n + t_n b'_n) - g_n(b_n) - \nabla g_n(b_n, b'_n) \\ &\quad + \nabla g_n(b_n, b'_n - \nabla f_n(a_\nabla, a_0)) \\ &= o(t_n) + o(t_n) = o(t_n) \end{aligned}$$

for every $t_n \to 0$ and every $a_n \to a_0$ with $a_\nabla + t_n a_n \in \mathbb{A}_f \; \forall n \in \mathbb{N}$. Similar reasoning applies to uniform equi-differentiability of the second and third kinds. □

Let us finally observe that the concept of *Uniform Hadamard Equi-Differentiability* is amenable to the product topology 'treatment' that we have explored in Propositions C.11, C.11 and C.31.

Proposition C.40. (Uniform Hadamard Equi-Differentiability with Product Topology) *Let* $(\mathbb{A}, \mathcal{T}_\mathbb{A})$ *and* $(\mathbb{B}_i, \mathcal{T}_{\mathbb{B}_i})$ *be topological vector spaces for every* i *in some countable index set* $\mathbb{I}$ *and* $(\mathbb{B}, \mathcal{T}_\mathbb{B})$ *be the product space* $\mathbb{B} = \times_{i \in \mathbb{I}} \mathbb{B}_i$ *with product topology* $\mathcal{T}_\mathbb{B}$. *Let* $\{f_n\}_{n \in \mathbb{N}}$ *of Hadamard differentiable functions* $f_n : \mathbb{A}_f \to \mathbb{B} \;\; \forall n \in \mathbb{N}$ *with derivative at* a_∇ *denoted* $\nabla_{\mathbb{A}_0} f_n(a_\nabla, \cdot) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$. *Then, a sequence of functions*

$\{f_n\}_{n\in\mathbb{N}}$ with $f_n : \mathbb{A}_f \to \mathbb{B}\ \ \forall n \in \mathbb{N}$, is uniform Hadamard equi-differentiable of the first kind, second or third kinds, along every sequence $a_n \to a_\nabla \in \mathbb{A}_\nabla$, tangentially to $\mathbb{A}_0$, if and only if the coordinate projections $\pi_i f_n : \mathbb{A}_f \to \mathbb{B}_i$ are uniform Hadamard equi-differentiable of the first, second or third kinds respectively, along every sequence $a_n \to a_\nabla \in \mathbb{A}_\nabla$, tangentially to $\mathbb{A}_0$, for every $i \in \mathbb{I}$.

Proof. By definition, a sequence $\{f_n\}_{n\in\mathbb{N}}$ of Hadamard differentiable functions $f_n : \mathbb{A}_f \to \mathbb{B}\ \ \forall n \in \mathbb{N}$ with derivative at a_∇ denoted $\nabla_{\mathbb{A}_0} f_n(a_\nabla, \cdot) \in \mathbb{L}(\mathbb{A}_0, \mathbb{B})$ is said to be uniform Hadamard equi-differentiable of the first kind, along every sequence $a_n \to a_\nabla \in \mathbb{A}_\nabla$, tangentially to $\mathbb{A}_0$, if the sequence $\left\{b_n(t_n, a_n, a'_n)\right\}_{T\in\mathbb{N}} \subset \mathbb{B}$ defined as,

$$b_n(t_n, a_n, a'_n) := \frac{f_n(a_n + t_n a'_n) - f_n(a_n) - \nabla f_n(a_\nabla, a_0)}{t_n}$$

converges to zero, for every $t_n \to 0$, every $a_n \to a_\nabla$, and every $a'_n \to a_0$ with $a_n + t_n a'_n \in \mathbb{A}_f\ \forall n \in \mathbb{N}$. Now, by Corollary A.16 given the product topology $\mathcal{T}_\mathbb{B}$ on $\mathbb{B}$, convergence of the sequence $b_n(t_n, a_n, a'_n) \to 0$ on the product space $\mathbb{B}$ occurs if and only if its coordinate projections $\pi_i b_n(t_n, a_n, a'_n)$ vanish $\pi_i b_n(t_n, a_n, a'_n) \to 0$ in $\mathbb{B}_i$ for every $i \in \mathbb{I}$. By linearity and continuity of the coordinate projection (Lemma A.15 and Proposition A.50) and the definition of Hadamard derivative, it follows immediately that $\nabla\pi_i(\boldsymbol{\beta}) = \pi_i \in \mathbb{L}(\mathbb{A}, \mathbb{B})$ for every $i \in \mathbb{I}$. By the chain rule we thus obtain,

$$\nabla_{\mathbb{A}_0} \pi_i f_n(a_\nabla, a_0) = \pi_i \nabla_{\mathbb{A}_0} f_n(a_\nabla, a_0).$$

As a result, we then have that $b_n(t_n, a_n, a'_n) \to 0$ if and only if,

$$\begin{aligned}\pi_i b_n(t_n, a_n, a'_n) : &= \pi_i \left(\frac{f_n(a_n + t_n a'_n) - f_n(a_n) - t_n \nabla f_n(a_\nabla, a_0)}{t_n} \right) \\ &= \frac{\pi_i f_n(a_n + t_n a'_n) - \pi_i f_n(a_n) - t_n \nabla \pi_i f_n(a_\nabla, a_0)}{t_n} \to 0 \text{ for every } i \in \mathbb{I}.\end{aligned} \tag{C.7}$$

Finally, since by Lemma B.5 a composition of linear maps is linear, and by lemma A.29 a composition of continuous maps is continuous, $\pi_i \circ \nabla_{\mathbb{A}_0} f(a_\nabla)$ is a continuous linear map on $\mathbb{A}_0$. This implies, by definition, that the convergence in (C.7) above holds if and only if $\pi_i f$ is uniformly Hadamard differentiable of the first kind along every sequence $a_n \to a_\nabla$ for every $i \in \mathbb{I}$. The same argument applies to sequences, $b_n(t_n, a_n, a'_n) := t_n^{-1}(f_n(a_n + t_n a'_n) - f_n(a_n) - t_n \nabla_{\mathbb{A}_0} f_n(a_n, a_0))$ and $b_n(t_n, a_n, a'_n) := t_n^{-1}(f_n(a_n + t_n a'_n) - f_n(a_n) - t_n \nabla_{\mathbb{A}_0} f_n(a_n, a'_n))$, and hence, also to uniform Hadamard differentiability of the second and third kinds. □

Appendix D

Normalization of Variables in Simulations from Dynamic Models

Dynamic models derived from economic theory often deal with variables defined in measurement units that do not correspond to those of observed data. This reflects the fact that the absolute magnitude of economic data is usually meaningless and that economists are instead interested in relative properties of these variables and the relations between them. However, the linear approximation of functions establishing these relations may not be robust to changes in unit of measurement. Hence, normalization procedures that are mean-shifting can have important effects that should not be ignored. This is especially important when comparing alternative solution methods to DSGE models since normalizing constants may unintentionally enhance the benefits of nonlinear solution methods or obscure the deficiencies of linear ones. The normalization problem is always present since even the abssentist researcher is unwillingly imposing one. In applied work, from the first RBC models of Kydland and Prescott (1982) to today's DSGEs of e.g. Christiano et al. (2005), it seems that their has been widespread lack of attention given to this subject. The literature dealing with nonlinear solution methods for rational expectation models, from Taylor and Uhlig (1990) and Judd (1992) to Aruoba et al. (2006), seems to have also ignored this point when conducting simulation based exercises. The magnitude of these effects is likely to depend on the nature of the functions being approximated, the properties and the size of the dynamic model itself.

D.1 Normalization and Linear Approximation

Let $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^k$ denote a k-dimensional vector random variable and $\{\mathbf{x}_t\}_{t=1}^{\infty}$ a stationary ergodic stochastic sequence whose time-invariant conditional density is implicitly defined by the dynamic stochastic model $\mathbf{x}_{t+1} = f(\mathbf{x}_t) + \epsilon_t$ where, for

instance, ϵ_t is an iid process that follows some distribution G. Furthermore, suppose that we are interested in linearizing $f : \mathcal{X} \to \mathcal{X}$ about $\mathbf{x}_0$, here taken to be the mean of $\mathbf{x}_t$ (i.e. $\mathbf{x}_0 \equiv E(\mathbf{x}_t)$), although it could also be some unique time-invariant steady-state of a non-stochastic related model.[1] Then assuming that $f \in \mathrm{C}^2(\mathcal{X})$, the space of twice continuously differentiable functions on $\mathcal{X}$, we have by Taylor's theorem that $f(\mathbf{x}_t) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x}_t - \mathbf{x}_0) + R_2(\mathbf{x}_t)$ where $R_2(\mathbf{x}_t) = \sum_{|\alpha|=2}(2/\alpha!)\left[\int_0^1 (1-t)D^\alpha f(\mathbf{x}_0 + t(\mathbf{x}_t - \mathbf{x}_0))dt\right](\mathbf{x}_t - \mathbf{x}_0)^\alpha$, in multi-index notation. The form of the remainder function is of special importance to us since it reveals that R_2 is also a function of $\mathbf{x}_0$. In particular, Taylor's theorem shows that, when linearizing functions, economists should be concerned with (*i*) the size of the fluctuations of $\mathbf{x}_t$ around $\mathbf{x}_0$, (*ii*) the curvature of f, i.e. the magnitude of second derivatives $D^\alpha f$ on $\mathcal{X}$, and (*iii*) the implications of normalization procedures that shift $\mathbf{x}_0 \in \mathcal{X}$. While (*i*) and (*ii*) are generally well understood, (*iii*) is often ignored. This chapter of the Appendix is precisely devoted to addressing the nature of (*iii*) and describing the adverse practical consequences of ignoring it. In this respect, we note that the problem must be framed in terms of a stochastic sequence whose variation is in some sense "small", since otherwise approximations errors could easily be unbounded.[2] Fortunately, observation of a sample path $\{\mathbf{x}_t\}_{t=1}^T \in \mathcal{X}^T$ gives information to the researcher about the variability of $\mathbf{x}_t$ around $\mathbf{x}_0$, usually summarized by measures such as the sample variance. The researcher is then interested in having a small approximation error, at least within some interval $[\delta\mathbf{x}_0, (1+\delta)\mathbf{x}_0] \subseteq \mathcal{X}$ and the approximation error can be deemed "acceptable" if for some constant $\epsilon > 0$, we have $|R_2(\mathbf{x}_t)| < \epsilon \,\forall (1-\delta)\mathbf{x}_0 \leq \mathbf{x}_t \leq (1+\delta)\mathbf{x}_0$. As shown above, this error might not be invariant w.r.t. $\mathbf{x}_0$. In practice, one simple way of analyzing this is also to look at the linear approximation errors at the boundaries of the interval, which are given by $R_2((1-\delta)\mathbf{x}_0) = f((1-\delta)\mathbf{x}_0) - \sum_{|\alpha|=0}^{|\alpha|=1} \frac{D^\alpha f(\mathbf{x}_0)}{\alpha!}((1-\delta)\mathbf{x}_0)^\alpha$, and $R_2((1+\delta)\mathbf{x}_0) = f((1+\delta)\mathbf{x}_0) - \sum_{|\alpha|=0}^{|\alpha|=1} \frac{D^\alpha f(\mathbf{x}_0)}{\alpha!}((1+\delta)\mathbf{x}_0)^\alpha$.

D.2 Normalization and the Absolute Error of Linear Approximation

For simplicity, suppose that $\dim(\mathbf{x}_t) = 1$, and let $T_f(x_t, x_0)$ denote the linear approximation of f around x_0 evaluated at x_t, i.e. $T_f(x_t, x_0) := f(x_0) + f'(x_0)(x_t - x_0)$. The error of approximation at the bounds $(1-\delta)x_0$ and $(1+\delta)x_0$, can be used to check for

[1] Setting $\mathbf{x}_0 \equiv E(\mathbf{x}_t)$, $|E(\mathbf{x}_t)| < \infty$ is not only a natural choice but also one that is justified by the objective of minimizing the expected value of the linear approximation error R_2 i.e. $\mathbf{x}_0 = \arg\min_{\mathbf{x}_0 \in \mathcal{X}} E(R_2) \Rightarrow \mathbf{x}_0 = E(\mathbf{x}_t)$.

[2] This can in principle be made precise by requiring, for instance, that for some bounded $\mathcal{X}^* \subseteq \mathcal{X}$ we have $\mathbf{x}_t \in \mathcal{X}^* \;\forall t$ with probability one.

error invariance w.r.t. normalization procedures. In particular, $f(x_t) - T_f(x_t, x_0) = g(x_t, x_0)$ is invariant w.r.t. x_0 iff $g(x_t, x_0) = g(x_t)$ (not a function of x_0). For concreteness, we now apply this reasoning to functions that provide simple and intuitive illustrations of the behavior of absolute linear approximation errors. Define the linear approximation errors e_M at the maximum bound by,[3]

$$e_M(x_0, \delta) := f\Big(x_0(1+\delta)\Big) - \Big[f(x_0) + \frac{\partial f}{\partial x}\Big|_{x=x_0}\Big(x_0(1+\delta) - x_0\Big)\Big].$$

Clearly, on a function like $f(x_t) = \log(x_t)$, these errors are invariant to normalizations that shift x_0 since,

$$\begin{aligned} e_M(x_0, \delta) &= \log\Big(x_0(1+\delta)\Big) - \left[\log(x_0) + \frac{1}{x_0}\Big(x_0(1+\delta) - x_0\Big)\right] \\ &= -\delta + \log(1+\delta) = e_M(\delta), \end{aligned}$$

is not a function of x_0 (which also holds for the minimum bound e_m). The same result can be found by looking at the following form of the Taylor's expansion nth order remainder

$$\begin{aligned} & R_n(x_0, x_t, \xi) \equiv \frac{1}{(n+1)!}\frac{\partial^{n+1} f(\xi)}{\partial \xi^{n+1}}(x_t - x_0)^{n+1} \\ \Leftrightarrow\ & R_n(x_0, \delta, \delta_\xi) = \frac{1}{(n+1)!}\frac{\partial^{n+1} f(\delta_\xi x_0)}{\partial(\delta_\xi x_0)^{n+1}}(\delta x_0)^{n+1}, \end{aligned}$$

where $\xi \in (x_0, x_t) \vee (x_t, x_0)$, and on the r.h.s., the remainder is written in terms of a δ deviation from the steady-state x_0, by defining $\delta_\xi \equiv \xi/x_0$ and $\delta = (x_t - x_0)/x_0$. Now, verifying that the remainder is invariant to x_0 for $f(x_t) = \log(x_t)$ is immediate since,

$$\frac{\partial^2 f(\delta_\xi x_0)}{\partial(\delta_\xi x_0)^2} = -\frac{1}{(\delta_\xi x_0)^2} \Rightarrow R_2(x_0, \delta, \delta_\xi) = -\frac{1}{2(\delta_\xi x_0)^2}(\delta x_0)^2 = -\frac{\delta^2}{2\delta_\xi^2} = R_2(\delta, \delta_\xi).$$

However, as we shall now see, this is not the case for those functions typically featured in DSGE models.

D.2.1 Cobb-Douglas and CRRA Functions

The sensitivity of linear approximation errors w.r.t. normalization procedures is present in commonly used functions such as the Cobb-Douglas $f(k_t, h_t) = k_t^\alpha h_t^{1-\alpha}$ and the CRRA utility function $u(c_t) = \frac{c_t^{1-\theta}}{1-\theta}$. Indeed, in the case of $u(c_t)$, the approximation errors are not invariant to normalization of its arguments since,

$$R_2(x_0, \delta, \delta_\xi) = -\frac{1}{2}\theta(\delta_\xi x_0)^{-\theta-1}(\delta x_0)^2 = -\frac{1}{2}\theta\delta_\xi^{-\theta-1}\delta^2 x_0^{1-\theta},$$

[3]Substituting $(1+\delta)$ by $(1-\delta)$ yields the approximation error $e_m(x_0, \delta)$ at the lower bound.

is invariant to changes in x_0 only when $\theta = 1$ (which takes as a limit case the $\log(x_t)$) and furthermore,

$$0 < \theta < 1 \;\Rightarrow\; \frac{\partial R_2(x_0, \delta, \delta_\xi)}{\partial x_0} > 0\,, \quad \theta > 1 \;\Rightarrow\; \frac{\partial R_2(x_0, \delta, \delta_\xi)}{\partial x_0} < 0\,.$$

Figure D.1 plots the error of approximation of the CRRA utility function for $\theta = 0.5$ (left) and $\theta = 2$ (right), with $c_t = \delta c_0$, $\delta \in [-0.1, 0.1]$. Clearly, the opposite behavior of approximation errors for $\theta < 1$ and $\theta > 1$ makes it hard to produce general advice to economists that is always valid.

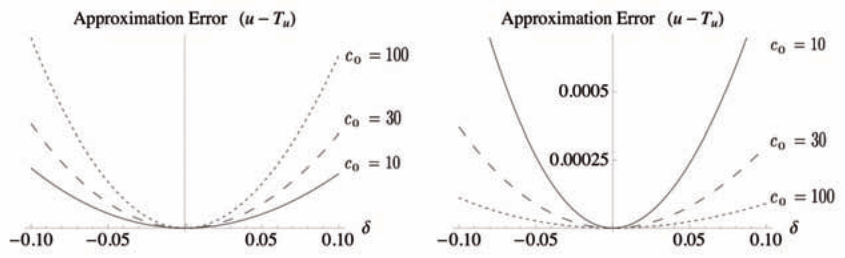

Figure D.1: Approximation errors of $u(c_t)$, for $\theta = 0.5$ (left) and $\theta = 2$ (right).

Consider now the Cobb-Douglas production function, $f(k_t, h_t) = k_t^\alpha h_t^{1-\alpha}$. The linear approximation error of a (δ_k, δ_h) relative deviation from the expansion point (k_0, h_0) is given by,

$$e_M(k_0, h_0, \delta_k, \delta_h) = k_0^\alpha h_0^{1-\alpha}\Big[(1+\delta_k)^\alpha (1+\delta_h)^{1-\alpha} - \alpha\delta_k - (1-\alpha)\delta_h - 1\Big],$$

which, as Figure D.2 reveals, for $k_t = \delta_k k_0$, $\delta_k \in [-0.1, 0.1]$ and $h_t = \delta_h h_0$, $\delta_h \in [-0.1, 0.1]$ is increasing in both dimensions.

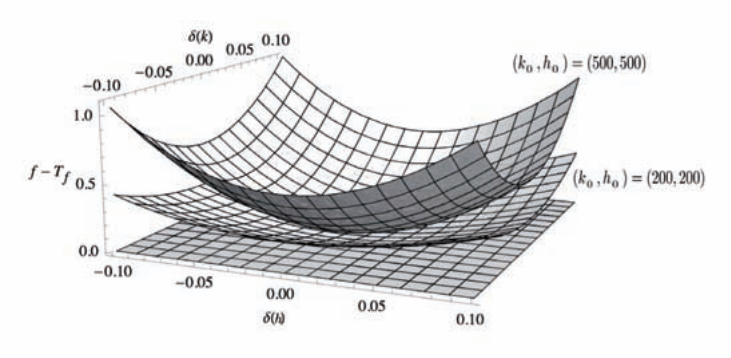

Figure D.2: Linear approximation errors of Cobb-Douglas production function $f(k_t, h_t) = k_t^\alpha h_t^{1-\alpha}$ about (k_0, h_0).

D.3 Normalization and the Relative Error of Approximation

One important feature of the Cobb-Douglas and CRRA functions encountered above is their homogeneity. The importance of this property stems from the fact that the relative error of approximation of homogeneous functions is always constant, i.e. even though for any given δ,

$$u(\delta c_0) - T_u(\delta c_0; c_0) \quad \text{and} \quad f(\delta k_0, \delta z_0) - T_f\Big((\delta k_0, \delta z_0); (k_0, z_0)\Big)$$

may change with the choice of c_0 and (k_0, z_0) respectively, we always have that $\Big[u(\delta c_0) - T_u(\delta c_0; c_0)\Big]/u(c_0)$ and $\Big[f(\delta k_0, \delta z_0) - T_f\Big((\delta k_0, \delta z_0); (k_0, z_0)\Big)\Big]/f(k_0, z_0)$ are invariant to normalization. This is important because the relative error of approximation is often more interesting than the absolute one. Consider for instance a production function $y_t = f(x_t)$. Knowing that the error approximation as a fraction of output y_t is invariant to shifts in x_0 might be satisfactory enough in many applications.

Lemma D.3.1. (Homogeneous Function Invariance) *Let $f(\phi x_t) = \phi^\nu f(x_t)$ be C^2 in $\mathcal{X}$ and define $e^r_M(x_0, \delta) := e_M(x_0, \delta)/f(x_0)$, $R^r_2(x_0, \delta, \delta_\xi) := R_2(x_0, \delta, \delta_\xi)/f(x_0)$. Then, for $x_t = (1+\delta)x_0 \in \mathcal{X} \subseteq \mathbb{R}$, $x_0 \in \text{int}(\mathcal{X})$ and $\delta_\xi \in (1, 1+\delta) \vee (1+\delta, 1)$, we have $e^r_M(x_0, \delta) = (1+\delta)^\nu - (1+\nu\delta) = e^r_M(\delta)$ and*

$$R^r_2(x_0, \delta, \delta_\xi) = (\nu^2 - \nu)\frac{\delta^2}{2\delta_\xi^{2-\nu}} = R^r_2(\delta, \delta_\xi),$$

i.e. the relative linear approximation error is invariant to mean-shifting normalization.

As we shall see now, this does not mean, however, that economists should not be concerned with the normalization of variables in dynamic models, even when they have in mind only relative approximation errors.

D.4 Normalization and Dynamic Models

While homogeneous functions share the interesting invariance property described above, we must be reminded of the fact that difference equations commonly found in structural models, e.g. DSGE models, establish non-homogeneous functional relations between variables at time t and their past. For example, the relative error of a linear approximation of a simple capital accumulation process, $k_{t+1} = (1-\rho)k_t + f(k_t)$ will in general depend on the steady-state of capital k_0 because, even if $f(k_t)$

is homogeneous, $F(k_t) = (1-\rho)k_t + f(k_t)$ is not. Indeed, for $k_{t+1} = (1-\rho)k_t + k_t^\alpha$,

$$R_2^r(x_0, \delta, \delta_c) = \alpha(\alpha-1)\frac{\delta^2}{2\delta_\xi^{2-\alpha}} k_0^\alpha \Big((1-\rho)k_0 + k_0^\alpha\Big)^{-1}.$$

Figure D.3 plots the relative linear approximation error of $k_{t+1} = F(k_t) = (1-\rho)k_t + k_t^\alpha$ for $\rho = 0.1$ and $\alpha = 0.5$ around k_0 for $k_t = \delta k_0$, $\delta \in [-0.1, 0.1]$. Note that, in persistent processes, the effect of these may accumulate over time to produce considerable effects on simulated paths.

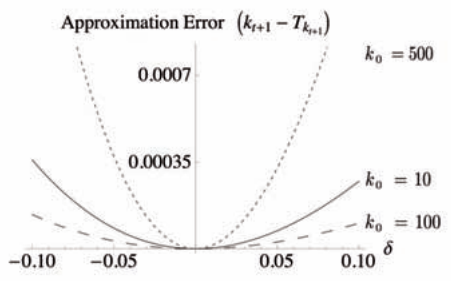

Figure D.3: Linear approximation errors of $k_{t+1} = F(k_t) = (1-\rho)k_t + k_t^\alpha$ about k_0.

In general, difference equations in DSGE models are simple composite functions where homogeneous functions are summable on the composition i.e. $F(f_1, .., f_{n_f}) = \sum_{j=1}^{n_f} f_i$ where $f_i(\phi \mathbf{x}_t) = \phi^{\nu_i} f_i(\mathbf{x}_t)$. Lemma 2 below, reveals that, when it comes to approximation error sensitivity to normalization, the different orders of homogeneity of these functions are typically to blame.

Lemma D.4.1. (Composition Invariance) *Let* $F(\mathbf{x}_t) = \sum_{j=1}^{n_f} f_i(\mathbf{x}_t)$ *with* $f_i(\phi \mathbf{x}_t) = \phi^{\nu_i} f_i(\mathbf{x}_t)$. *Then, for* $\mathbf{x}_t = (1+\delta)\mathbf{x}_0 \in \mathcal{X} \subseteq \mathbb{R}^k$, $\mathbf{x}_0 \in \text{int}(\mathcal{X})$ *and* $\delta_\xi \in (1, 1+\delta) \vee \delta_\xi \in (1+\delta, 1)$, *the relative linear approximation error* $F(\delta \mathbf{x}_0) - T_F(\delta \mathbf{x}_0)$ *is invariant to* $\mathbf{x}_0$ *if* $\nu_i = \nu \, \forall i$.

In the example above, relative linearization errors are not invariant to normalization because $(1-\rho)k_t$ and k_t^α are homogeneous functions of different degrees. This result is generally applicable to nonlinear dynamic structural models in economics since typically these models are defined by difference equations $F(\mathbf{x}_t) = \sum_{j=1}^{n_f} f_i(\mathbf{x}_t)$ with $f_i(\phi \mathbf{x}_t) = \phi^{\nu_i} f_i(x)$ for which $\exists (i,j) : \nu_i \neq \nu_j$ thus making the composition non-homogeneous.

References

Ahuja, G. C., Narang, T. D., and Trehan, S. (1977). Best approximation on convex sets in metric linear spaces. *Mathematische Nachrichten*, 78:125–130.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2 nd, Tsahkadsor, Armenian SSR*, pages 267–281.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1):3–14.

Amemiya, T. (1983). Non-linear regression models. *Handbook of Econometrics*, 1:333–389.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Andrews, D. W. (1986). Empirical process methods in econometrics. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, chapter 37, pages 2247–2294. Elsevier.

Andrews, D. W. (1992). Generic uniform convergence. *Econometric Theory*, 8:241–257.

Andrews, D. W. K. (1984). Non-strong mixing autoregressive processes. *Journal of Applied Probability*, 21(4).

Andrews, D. W. K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers [on unification of the asymptotic theory of nonlinear econometric models]. *Econometrica*, 55(6):1465–71.

Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica*, 59(2):307–45.

Angelos, J. and Egger, A. (1984). Strong uniqueness in L^p spaces. *Journal of Approximation Theory*, 42:14–26.

Aruoba, S., Fernandez-Villaverde, J., and Rubio-Ramirez, F. (2006). Comparing solution methods for dynamic equilibrium economies. *Journal of Economic Dynamics and Control*, 30(12):2477–2508.

Basener, W. F. (1973). *Topology and its Applications*. Pure and Applied Mathematics: A Wiley-Interscience. John Wiley and Sons, Inc.

Bates, C. and White, H. (1985). A unified theory of consistent estimation for parametric models. *Econometric Theory*, 1(02):151–178.

Bergstrom, A. R. (1985). Non-parametric functions in a Hilbert space. *Econometric Theory*, 1:7–26.

Billingsley, P. (1995). *Probability and Measure*. Wiley-Interscience.

Billio, M. and Monfort, A. (2003). Kernel-based indirect inference. *Journal of Financial Econometrics*, 1(3):297–326.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150. 10.1007/BF01199316.

Birge, L. and Massart, P. (1998). Minimum constrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli*, (4):329–375.

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Annals of Statistics*, 1:902–912.

Burguete, J., Gallant, A. R., and Souza, G. (1982). On unification of the asymptotic theory of nonlinear econometric models. *Econometric Reviews*, 1(2):151–190.

Chebana, F. (2007). M-processes and applications. *C. R. Math.*, 344(4):265–270.

Chebana, F. (2009). Parametric estimation with a class of M-estimators. *Mathematical Methods of Statistics*, 18(3):231–140.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, volume 6 of *Handbook of Econometrics*, chapter 76. Elsevier.

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, (71):1591–1608.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, (66):289–314.

Chen, X. and White, H. (1998). Nonparametric adaptive learning with feedback. *Journal of Economic Theory*, 82(1):190 – 222.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45:682–691.

Cheney, E. (1982). *Approximation Theory*. American Mathematical Society Chelsea Publishing, 2nd edition.

Cheney, E. W. (1974). Letter to the editor: Best approximation on convex sets in a metric space. *Journal of Approximation Theory*, 12:94–97.

Christiano, L., Eichenbaum, M., and Evans, C. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *The Journal of Political Economy*, 113(1):1–45.

Clarke, B. (1983). Uniqueness and Frechet differentiability of functional solutions to maximum likelihood. *The Annals of Statistics*, 11(4):1196–1205.

Clarkson, J. A. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton UNiverstity Press.

Crisp, A. and Burridge, J. (1993). A note on the uniqueness of M-etimators in robust regression. *The Canadian Journal of Statistics*, 21(2):205–208.

Dave, C. and Dejong, D. N. (2007). *Structural Macroeconometrics*. Princeton University Press.

Davidson, J. (1994). *Stochastic Limit Theory*. Advanced Texts in Econometrics. Oxford University Press.

Davidson, K. R. and Donsig, A. P. (2009). *Real Analysis and Applications: Theory in Practice*. Springer, 1 edition.

de Boor, C. (1978). *A practical guide to splines*. Springer-Verlag, New York.

Debreu (1967). Integration of correspondences. *Proceeding of the Fifth Berkley Symposium on Mathematical Statistics and Probability*, 2(1).

Debreu, G. (1959). Theory of value : An axiomatic analysis of economic equilibrium. *Monograph (Yale University. Cowles Foundation for research in Economics)*, (17).

Denkowski, Z., Migorski, S., and Papageorgiou, N. S. (2003). *An Introduction to Nonlinear Analysis: Theory*. Kluwer Academic/Plenum Publishers, New York.

Dhaene, G., Gourieroux, C., and Scaillet, O. (1998). Instrumental models and indirect encompassing. *Econometrica*, 66(3):673–688.

Diewert, W. E. and Wales, T. J. (1987). Flexible functional forms and global curvature conditions. *Econometrica*, 55(1):43–68.

Domowitz, I. (1985). New directions in non-linear estimation with dependent observations. *Canadian Journal of Economics*, 18(1):1–27.

Domowitz, I. and White, H. (1982). Misspecified models with dependent observations. *Journal of Econometrics*, 20(1):35–58.

Donoho, D. L. and Liu, R. C. (1988). Pathologies of some minimum distance estimators. *The Annals of Statistics*, 16(2):587–608.

Doob, J. L. (1934). Probability and statistics. *Transactions of the American Mathematical Society*, 36(4):759–775.

Doob, J. L. (1953). *Stochastic Processes*. John Wiley and Sons.

Dridi, R. and Renault, E. (2000). Semi-parametric indirect inference. STICERD - Econometrics Paper Series /2000/392, Suntory and Toyota International Centres for Economics and Related Disciplines, LSE.

Ducharme, G. R. (1995). Uniqueness of the least-distances estimator in regression models with multivariate response. *The Canadian Journal of Statistics*, 23(4):421–424.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Foland, G. B. (2009). *A Guide to Advanced Real Analysis*. Dolciani Mathematical Expositions. Cambridge University Press.

Freedman, D. A. and Diaconis, P. (1982). On inconsistent M-estimators. *The Annals of Statistics*, 10(2):454–461.

Gallant, A. R. (1975). Nonlinear regression. *The American Statistician*, 29(2):73–81.

Gallant, A. R. (1981). On the bias in flexible functional forms and an essentially unbiased form: The Fourier flexible form. *Journal of Econometrics*, 15:211–245.

Gallant, A. R. (1986). *Nonlinear statistical models*. John Wiley & Sons, Inc., New York, NY, USA.

Gallant, A. R. (1987). Identification and consistency in seminonparametric regression. *Bewley, Truman F., ed. (1987), Advances in Econometrics Fifth World Congress*, 1:145–170.

Gallant, A. R. and Golub, G. H. (1984). Imposing curvature restrictions on flexible functional forms. *Journal of Econometrics*, 26(3):295–321.

Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55:363–390.

Gallant, A. R. and Souza, G. (1991). On the asymptotic normality of fourier flexible form estimates. *Journal of Econometrics*, 50(3):329–353.

Gallant, A. R. and White, H. (1988a). There exists a neural network that does not make avoidable mistakes. In *International Symposium on Neural Networks*.

Gallant, A. R. and White, H. (1992). On learning the derivatives of an unknown mapping with multilayer feedforward networks. *Neural Networks*, 5:129–138.

Gallant, R. and White, H. (1988b). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Cambridge University Press.

Gamelin, T. W. and Greene, R. E. (1999). *Introduction to Topology*. Dover Publications.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10(2):401–414.

Genton, M. G. and Ronchetti, E. (2003). Robust indirect inference. *Journal of the American Statistical Association*, 98(461):67–76.

Gill, R. D. (1986). Non- and semi-parametric maximum likelihood estimators and the von Mises method (part 1. Technical report, Centrum voor Wiskunde en Informatica.

Gouriéroux, C. and Monfort, A. (1993). Encompassing and indirect inference. *Statistical Methods amp; Applications*, 2:291–307. 10.1007/BF02589066.

Gourieroux, C. and Monfort, A. (1996). *Simulation-based Econometric Methods*. CORE Lectures. Oxford University Press.

Gourieroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:85–118.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3):681–700.

Granger, C. W. J. and Terasvirta, T. (1993). *Modelling Non-Linear Economic Relationships*. Number 9780198773207 in OUP Catalogue. Oxford University Press.

Grenander, U. (1981). *Abstract inference*. Wiley, New York :.

Gusak, D., Kukush, A., Kulik, A., Mishura, Y., and Pilipenko, A. (2010). *Theory of stochastic processes. With applications to financial mathematics and risk theory.* Problem Books in Mathematics. New York, NY: Springer. xii, 375 p. EUR 64.15 .

Han, C. and Phillips, P. C. B. (2006). GMM with many moment conditions. *Econometrica*, 74(1):147–192.

Hannan, E. (1970). Non-linear time series regression. Cowles Foundation Discussion Papers 298, Cowles Foundation for Research in Economics, Yale University.

Hanner, O. (1956). On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3(19):239–244.

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. In *Proceedings of IEEE First Annual International Conference on Neural Networks*, volume 3, pages III–11–III–14.

Hildenbrand, W. (1974). *Core and Equilibria of a Large Economy*. Princeton University Press.

Hornik, K., Stinchcombe, M., White, H., and Auer, P. (1994). Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Comput.*, 6:1262–1275.

Hornik, K., Stinchcombe, M. B., and White, H. (1989). Multilayer feedfoward networks are universal approximators. *Neural Networks*, 2:359–366.

Hsiao, C. (1983). *Handbook of Econometrics*, volume 1, chapter Identification. North Holland Publishing Company.

Huang, J. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, (28):960–999.

Huber, P. (1974). *Robust Statistics*. Wiley, New York.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the berkeley symposium on mathematical statistics an*.

James, I. M. (1987). *Topological and Uniform Spaces*. Springer-Verlag, New York Inc.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Judd, K. (1992). Projection methods for solving aggregate growth models. *Journal of Economic Theory*, 58(2):410–452.

Judd, K. (1998). Numerical methods in economics. *Cambridge, Mass: MIT Press*.

Kabaila, P. (1983). Parameter values of ARMA models minimising the one-step-ahead prediction error when the true system is not in the model set. *Journal of Applied Probability*, 20(2):405–408.

Kent, J. T. and Tyler, D. E. (2001). Regularity and uniqueness for constrained M-estimates and redescending M-estimates. *The Annals of Statistics*, 29(1):252–265.

Klambauer, G. (1973). *Real analysis*. North-Holland, New York, NY.

Klein, E. and Thompson, A. (1984). *Theory of correspondences: including applications to mathematical economics*. Canadian Mathematical Society series of monographs and advanced texts. Wiley.

Kolmogorov, A. N. and Fomin, S. V. (1975). *Introductory Real Analysis*. Dover Publications.

Krantz, S. and Parks, H. (1992). *A Primer of Real Analytic Functions*. Birkhauser Advanced Texts, second edition edition.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kydland, F. E. and Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica*, 50(6):1345–1370.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Universtity of California Publications in Statistics*, 2:23–53.

Lee, J. M. (2000). *Introduction to Topological Manifolds*. Springer.

Leonard, I. E. and Sundaresan, K. (1974). Geometry of Lebesgue-Bochner function spaces-smoothness. *Transactions of the American Mathematical Society*, 198:229–251.

Lin, P. K. (1989). Strongly unique best approximation in uniformly convex banach spaces. *Journal of Approximation Theory*, 56:101–107.

Lucas, R. (1985). *Models of Business Cycles*. Basil Blackwell.

Luenberger, D. G. (1997). *Optimization by Vector Space Methods (Series in Decision and Control)*. Wiley-Interscience, 1969 edition.

Malinvaud, E. (1970). The consistency of nonlinear regressions. *The Annals of Mathematical Statistics*, 41(3):956–969.

Mann, H. B. and Wald, A. (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica*, 11(34).

McLeish, D. L. (1975). A maximal inequality and dependent strong laws. *Annals of Probability*, 3(5):829–839.

Monfort, A. (1996). A reappraisal of misspecified econometric models. *Econometric Theory*, 12(04):597–619.

Munkres, J. (2000). *Topology*. Prentice Hall, 2 edition.

Narang, T. D. (1981). Best approximation and strict convexity of metric spaces. *Archivum Mathematicum*, 017(2):87–90.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–67.

Newey, W. K. (1994). Series estimation of regression functionals. *Econometric Theory*, (10):1–28.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, (79):147–168.

Newman, D. J. and Shapiro, H. S. (1963). Some theorems on Chebyshev approximation. *Duke Mathematical Journal*, 30:673–681.

Nickl, R. and Pötscher, B. M. (2009). Efficient simulation-based minimum distance estimation and indirect inference. MPRA Paper 16608, University Library of Munich, Germany.

Nurberger, G. (1979). Unicity and strong unicity in approximation theory. *Journal of Approximation Theory*, 26:54–70.

Pollard, D. (1989). Asymptotics via empirical processes. *Statistical Science*, 4:341–354.

Pollard, D. (1990). *Empirical processes: theory and applications*. NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics.

Potscher, B. M. and Prucha, I. R. (1989). A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica*, 57(3):675–83.

Pötscher, B. M. and Prucha, I. R. (1991a). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part I: consistency and approximation concepts. *Econometric Reviews*, 10(2):125–216.

Pötscher, B. M. and Prucha, I. R. (1991b). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part II: Asymptotic normality. *Econometric Reviews*, 10(3):253–325.

Potscher, B. M. and Prucha, I. R. (1994). Generic uniform convergence and equicontinuity concepts for random functions : An exploration of the basic structure. *Journal of Econometrics*, 60(1-2):23–63.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models: Asymptotic Theory*. Springer-Verlag.

Potscher, B. M. and Prucha, I. R. (2001). *A Companion to Theoretical Econometrics*, chapter Basic Elements of Asymptotic Theory. Blackwell Publishing.

Powell, M. J. D. (1981). *Approximation Theory and Methods*. Cambridge University Press.

Ramsey, F. (1928). A mathematical theory of saving. *Economic Journal*, 38(152):543–559.

Reeds, J. A. (1976). On the definition of von Mises functionals. Research Report S-44, Dep. of Statistics, University of Harvard.

Reinsch, H. (1967). Smoothing by spline functions. *Numerische Mathematik*, 10:177–183.

Ren, J.-J. and Sen, P. K. (2001). Second order Hadamard differentiability in statistical applications. *Journal of Multivariate Analysis*, 77(2):187 – 228.

Rivest, L. P. (1989). De l'unicite des estimateurs robustes en regression lorsque le parametre d'echelle et le parametre de la regression sont estimes simultanement. *Canadian Journal of Statistics*, 17(2):141–153.

Robinson, P. M. (1972). Non-linear regression for multiple time-series. *Journal of Applied Probability*, 9(4):758–768.

Romaguera, S. and Sanchis, M. (2000). Semi-Lipschitz functions and best approximation in quasi-metric spaces. *Journal of Approximation Theory*, 103:292–301.

Ruge-Murcia, F. (2007). Methods to estimate dynamic stochastic general equilibrium models. *Journal of Economic Dynamics and Control*, 31(8):2599–2636.

S., C. and C., M. (2006). Best approximation in spaces with asymmetric norm. *Revue D'Analyse Numerique et de Theorie de L'approximation*, 35(1):17–31.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Mathematics*, 52:947–950.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. John Wiley & Sons, Inc.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics*, (25):2555–2591.

Shen, X. and Wong, W. (1994). Converge rate of sieve estimates. *The Annals of Statistics*, (22):580–615.

Smith, A. (1993). Estimating nonlinear time-series models using simulated vector autoregressions. *Journal of Applied Econometrics*, 8:563–584.

Stinchcombe, M. B. and White, H. (1992). Some measurability results for extrema of random functions over random sets. *Review of Economic Studies*, 59(3):495–514.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10(4):1040–1053.

Sundaresan, K. (1967). Smooth banach spaces. *Mathematische Annalen*, 173:191–199.

Sutherland, W. A. (2009). *Introduction to metric and topological spaces*. Oxford University Press, second edition edition.

Sviridyuk, G. A. and Fedorov, V. E. (2003). *Linear Sobolev type equations and degenerate semigroups of operators*. Inverse and Ill-Posed Problems Series. VSP.

Taylor, J. B. and Uhlig, H. (1990). Solving nonlinear stochastic growth models: A comparison of alternative solution methods. *Journal of Business and Economic Statistics*, 8(1):1–17.

Trapletti, A., Leisch, F., and Hornik, K. (1998). On the stationarity of autoregressive neural network models.

Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics*, 21:14–44.

Van de Geer, S. (1995). The method of sieves and minimum contrast estimators. *Mathematical Methods of Statistics*, (4):20–38.

van der Vaart, A. W. (1995). Efficiency of infinite dimensional M-estimators. *Statistica Neerlandica*, 49(1).

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer-Verlag, New York.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Institute of Mathematical Statistics*.

White, H. (1980a). Nonlinear regression on cross-section data. *Econometrica*, 48(3):721–46.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1):149–70.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

White, H. (1989a). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, 1:425–464.

White, H. (1989b). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association*, 84(408).

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary maps. *Neural Networks*, 3(535-549).

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge Books. Cambridge University Press.

White, H. and Domowitz, I. (1984). Nonlinear regression with dependent observations. *Econometrica*, 52(1):143–61.

White, H. and Wooldrige, J. M. (1991). Some results on sieve estimation with dependent observations. In Barnett, W., J., P., and G., T., editors, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge University Press.

Whittacker, E. T. (1923). On a new method of graduation. In *Proceedings of the Edinbourg Mathematical Society*, volume 41, pages 63–75.

Winitzki, S. (2010). *Linear Algebra via Exterior Products*. lulu.com, GNU Free Documentation License.

Wong, W. and Severini, T. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *The Annals of Statistics*, (19):603–632.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates for sieve MLE's. *The Annals of Statistics*, 23:339–362.

Wu, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Annals of Statistics*, 9(3):501–513.

Wulbert, D. E. (1971). Uniqueness and differential characterization of approximations from manifolds of functions. *American Journal of Mathematics*, 18:350–366.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorv's entropy. *Annals of Statistics*, 13(2):768–774.

Zhou, S., Shen, X., and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, (26):1760–1782.

Nederlandse Samenvatting

Dit proefschrift introduceert een nieuwe *zeefextremumschatter* die gebaseerd is op hulpstatistieken middels het principe van *indirecte inferentie*. Deze schatter is ontworpen als antwoord op twee gekende problemen in de econometrische analyse. Het eerste heeft betrekking op de restrictiviteit van parameterruimtes van eindige complexiteit, en de bijhorende restrictiviteit van correcte-specificatieaxioma's. Het tweede probleem behelst het mogelijke falen van klassieke schatters (e.g. door onhandelbare criteriumfuncties) in het geval van hoog-dimensionale dynamische modellen met niet-geobserveerde variabelen.

Dit proefschrift geeft primitieve condities voor de meetbaarheid, consistentie, convergentiesnelheid, en asymptotische verdeling van de *Semi-nietparametrische Indirecte-inferentieschatter*, ook SNPII-schatter genoemd. Meer bepaald, dit hoofdstuk leidt de consistentie, convergentiesnelheid en asymptotische *Gaussianiteit* af van SNPII-schatters die gebruik maken van een oneindig aantal parametrische hulpstatistieken. Gelijkaardige resultaten werden verkregen voor de schatting van gepaste functies van de ware parameter. Verder bevat het hoofdstuk een karakterisering van statistische inferentie, uitgevoerd met behulp van een dubbele benadering van de verdeling van de SNPII-schatter voor grote steekproeven.

Middels zogenaamde 'high-level'-aannames, introduceert dit eveneens nieuwe convergentiesnelheids- en asymptotische verdelingstheorema's die gelden voor de gehele klasse van *zeefextremumschatters* van gepaste 'smoothness'. Zulke algemene theorema's waren voorheen niet beschikbaar. Deze resultaten betekenen bijgevolg een uitbreiding op de bestaande literatuur over zeefextremumschatting.

Er wordt eveneens *Monte Carlo*-bewijs gegeven voor het gedrag van de SNPII-schatter in kleine steekproeven. Het proefschrift suggereert een aantal voordelen verbonden aan het gebruik van flexibele econometrische technieken, zoals SNPII, die algemeenheid verlenen aan correcte-specificatieaxioma's. Tevens wordt het gebruik van SNPII geanalyseerd in de context van theorie-gedreven modellen.

Tenslotte geeft het proefschrift aan dat de literatuur aangaande *benaderingstheorie* eveneens gebruikt kan worden om te verifiëren of *identificeerbare uniciteitsvoor-*

waarden gelden voor een klasse van extremumschatters op misgespecifieerde modellen. Meer bepaald, dit hoofdstuk reduceert de verificatie van *identificeerbare uniciteitscondities* tot het slechts verifiëren van *sterke uniciteit* van beste benaderingen. Het proefschrift biedt dus een theorie die *identificeerbare uniciteitsaannames* geeft die eenvoudiger te verifiëren zijn in verschillende contexten.

In conclusio, men kan zeggen dat dit proefschrift de eerste fundamentele resultaten bevat die een verdere ontwikkeling van *semi-nietparametrische indirecte-inferentieschatting* mogelijk maken.

Curriculum Vitae

Francisco Blasques was born on March 19, 1982 in Lisboa, Portugal. He completed a *Licenciatura* degree in Economics in 2006. He studied Economics at Maastricht University having obtained his Master's degree with distinction in 2008.

After graduation, Francisco joined the Department of Economics and the Department of Quantitative Economics as a Ph.D. candidate under the supervision of Prof. dr. Bertrand Candelon, Prof. dr. Jean-Pierre Urbain and co-supervision of dr. Eric Beutner. The results of his research are presented in this thesis.